Chapter 1
Probability Theory

1.1. Probability Space

Probability theory is concerned with measurements of random phenomena and the properties of such measurements. This opening section discusses the formulation of event structures, the axioms that need to be satisfied by a measurement to be a valid probability measure, and the basic set-theoretic properties of events and probability measures.

1.1.1. Events

At the outset we posit a set S, called the *sample space*, containing the possible experimental outcomes of interest. Mathematically, we simply postulate the existence of a set S to serve as a universe of discourse. Practically, all statements concerning the experiment must be framed in terms of elements in S and therefore S must be constrained relative to the experiment. Every physical outcome of the experiment should refer to a unique element of S. In effect, this practical constraint embodies two requirements: every physical outcome of the experiment must refer to some element in S and each physical outcome must refer to only one element in S. Elements of S are called *outcomes*.

Probability theory pertains to measures applied to subsets of a sample space. For mathematical reasons, subsets of interest must satisfy certain conditions. A collection $\mathcal{E}$ of subsets of a sample space S is called a *σ-algebra* if three conditions are satisfied:

(S1) $S \in \mathcal{E}$.

(S2) If $E \in \mathcal{E}$, then $E^c \in \mathcal{E}$, where E^c is the complement of E.

(S3) If the countable (possibly finite) collection $E_1, E_2, \ldots \in \mathcal{E}$, then the union $E_1 \cup E_2 \cup \cdots \in \mathcal{E}$.

Subsets of S that are elements of $\mathcal{E}$ are called *events*. If S is finite, then we usually take the set of all subsets of S to be the σ-algebra of events. However, when S is infinite, the problem is more delicate. An in-depth discussion of σ-algebras properly belongs to a course on measure theory and here we will restrict ourselves to a few basic points.

Since S is an event and the complement of an event is an event, the null set $\varnothing$ is an event. If $E_1, E_2, \ldots$ are events, then, according to De Morgan's law,

$$\bigcap_{i=1}^{\infty} E_i = \left(\bigcup_{i=1}^{\infty} E_i^c \right)^c \tag{1.1}$$

Since E_1^c, E_2^c,... are events, their union is an event, as is the complement of their union. Thus, a countable intersection of events is an event. In particular, set subtraction, $E_2 - E_1 = E_2 \cap E_1^c$, is an event.

On the real line $\Re$ there exists a smallest σ-algebra containing all open intervals (a, b), where $-\infty \leq a < b \leq \infty$. This σ-algebra, called the *Borel σ-algebra*, contains all intervals (open, closed, half-open-half-closed). It also contains all countable unions and intersections of intervals. For applications, we are only concerned with these countable unions and intersections. Sets in the Borel σ-algebra are called *Borel sets*.

Given a sample space S and a σ-algebra $\mathcal{E}$, a *probability measure* is a real-valued function P defined on the events in $\mathcal{E}$ such that the following three axioms are satisfied:

(P1) $P(E) \geq 0$ for any $E \in \mathcal{E}$.

(P2) $P(S) = 1$.

(P2) If $E_1, E_2,...$ is a disjoint (mutually exclusive) countable collection of events, then

$$P\left(\bigcup_{n=1}^{\infty} E_n \right) = \sum_{n=1}^{\infty} P(E_n) \tag{1.2}$$

The triple $(S, \mathcal{E}, P)$ is called a *probability space* and Eq. 1.2 is called the *countable additivity* property. In the case of two disjoint events, the additivity property reduces to

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) \tag{1.3}$$

***Example* 1.1.** If $S = \{a_1, a_2,..., a_n\}$ is a finite set of cardinality n, then a σ-algebra $\mathcal{E}$ is defined by the *power set* (set of all subsets) of S. For any singleton event $\{a_i\} \in \mathcal{E}$, assign a nonnegative value $P(\{a_i\})$ such that

$$\sum_{i=1}^{n} P(\{a_i\}) = 1$$

For any event $E = \{e_1, e_2,..., e_m\} \subset S$, define

$$P(E) = \sum_{j=1}^{m} P(\{e_j\})$$

and define $P(\varnothing) = 0$. Then P is a probability measure on $\mathcal{E}$. For convenience, the outcome probabilities $P(\{a_i\})$ are typically denoted by $P(a_i)$. In the special case where the outcomes are *equiprobable*, $P(a_i) = 1/n$ for $i = 1, 2,..., n$, event probabilities reduce to $P(E) = m/n$, where m is the cardinality of E. When a probability space is equiprobable, outcomes are said to occur *uniformly randomly*. ■

Example 1.2. An *urn model* consists of a set $U = \{1, 2,..., n\}$ representing n numbered balls in an urn and a protocol for randomly selecting balls from the urn. Probabilities are determined for events resulting from the protocol. We consider three protocols.

Ordered selection with replacement involves selecting $k > 0$ balls, one at a time, returning a selected ball to the urn, and recording in order the numbers of selected balls. A sample space for this protocol is given by the Cartesian product $U^k = U \times U \times \cdots \times U$, with each outcome being a k-vector. With uniformly random selection, the probability of any outcome $(b_1, b_2,..., b_k)$ is n^{-k} and event probabilities are thereby determined.

Ordered selection without replacement involves selecting $k \leq n$ balls without returning selected balls to the urn and recording in order the numbers of selected balls. Each outcome is a k-vector in which no component is repeated. Each such vector is known as a *permutation* of n objects taken k at a time. The number of such permutations is

$$P_{n,k} = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

When selecting k balls uniformly randomly with replacement, the probability of choosing a permutation (the event E composed of permutations) is

$$P(E) = \frac{n!}{n^k (n-k)!}$$

Unordered selection without replacement involves selecting $k \leq n$ balls without returning selected balls to the urn and recording without respect to order the numbers of selected balls. Each outcome is a subset of U and, in this context, is known as a *combination* of n objects taken k at a time. The number of such combinations is the number of subsets of U containing k elements and is given by

$$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The expression for $C_{n,k}$ follows from $P_{n,k}$ because for each subset containing k elements there are $k!$ permutations of the subset, so that $P_{n,k} = k!C_{n,k}$. ■

Both the counting of elements in a product set and of permutations are particular instances of a more general principle, namely, counting k-vectors formed according to the following scheme: (1) the first component of the vector can be occupied by any one of r_1 elements; (2) no matter which element is chosen for the first component, any one of r_2 elements can occupy the second component; (3) proceeding recursively, no matter which elements are chosen for the first $j - 1$ components, any one of r_j elements can occupy the jth component. According to the *fundamental principle of counting*, there are $r_1 r_2 \cdots r_k$ possible vectors that can result from application of the selection scheme.

***Example* 1.3.** Let $f(x)$ be a nonnegative integrable function defined on the real line $\mathfrak{R}$ whose integral over $\mathfrak{R}$ is unity. For any Borel set $B \subset \mathfrak{R}$, a probability measure on the Borel σ-algebra is defined by

$$P(B) = \int_B f(x)\,dx$$

The first two axioms of a probability measure are satisfied owing to the assumptions on f and the third is a fundamental property of integration. For instance, the two-event additivity property of Eq. 1.3 states that, if B_1 and B_2 are disjoint Borel sets, then

$$\int_{B_1 \cup B_2} f(x)\,dx = \int_{B_1} f(x)\,dx + \int_{B_2} f(x)\,dx \qquad \blacksquare$$

A number of basic properties of probability measures follow immediately from the axioms for a probability space $(S, \mathcal{E}, P)$. For any $E \in \mathcal{E}$, $E \cup E^c = S$. Applying additivity together with $P(S) = 1$ yields

$$P(E^c) = 1 - P(E) \tag{1.4}$$

It follows at once that

$$P(\varnothing) = P(S^c) = 1 - P(S) = 0 \tag{1.5}$$

If $E_1, E_2 \in \mathcal{E}$ and $E_1 \subset E_2$, then $E_2 = E_1 \cup (E_2 - E_1)$ and additivity implies

$$P(E_2 - E_1) = P(E_2) - P(E_1) \tag{1.6}$$

Since $P(E_2 - E_1)$ is nonnegative, $P(E_1) \leq P(E_2)$.

For any events $E_1, E_2 \in \mathcal{E}$, additivity implies

$$P(E_1 \cup E_2) = P(E_1 - E_2) + P(E_2 - E_1) + P(E_1 \cap E_2) \tag{1.7}$$

Using

$$P(E_1 - E_2) = P(E_1) - P(E_1 \cap E_2) \tag{1.8}$$

and an analogous expression for $P(E_2 - E_1)$, Eq. 1.7 becomes

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \tag{1.9}$$

Mathematical induction yields the *probability addition theorem*.

Theorem 1.1. If $(S, \mathcal{E}, P)$ is a probability space and $E_1, E_2, \ldots, E_n \in \mathcal{E}$, then

$$P\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{j=1}^{n} (-1)^{j+1} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} P\left(\bigcap_{k=1}^{j} E_{i_k}\right) \qquad \blacksquare \tag{1.10}$$

Theorem 1.2. If $(S, \mathcal{E}, P)$ is a probability space, $E_1, E_2, \ldots \in \mathcal{E}$, and $E_1 \subset E_2 \subset E_3 \subset \cdots$, then there is *continuity from below*:

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n \to \infty} P(E_n) \tag{1.11}$$

If $E_1 \supset E_2 \supset E_3 \supset \cdots$, then there is *continuity from above*:

$$P\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n \to \infty} P(E_n) \qquad \blacksquare \tag{1.12}$$

To show continuity from below, let $F_k = E_k - E_{k-1}$ for $k = 1, 2, \ldots$, where $E_0 = \varnothing$, and let E denote the union in Eq. 1.11. Then

$$E_n = \bigcup_{k=1}^{n} F_k \tag{1.13}$$

$$E = \bigcup_{k=1}^{\infty} F_k \tag{1.14}$$

where both unions are disjoint. Countable additivity applied to Eq. 1.13 yields

$$P(E_n) = \sum_{k=1}^{n} P(F_k) \tag{1.15}$$

Countable additivity applied to Eq. 1.14 yields

$$P(E) = \sum_{k=1}^{\infty} P(F_k) = \lim_{n\to\infty} \sum_{k=1}^{n} P(F_k) = \lim_{n\to\infty} P(E_n) \tag{1.16}$$

which verifies Eq. 1.11.

As for continuity from above, with $E_1 \supset E_2 \supset E_3 \supset \cdots$ and E denoting the intersection in Eq. 1.12, De Morgan's law together with continuity from below yields

$$\begin{aligned} P(E_1) - P(E) &= P\left(E_1 - \bigcap_{n=1}^{\infty} E_n\right) \\ &= P\left(\bigcup_{n=1}^{\infty} (E_1 - E_n)\right) \\ &= \lim_{n\to\infty} P(E_1 - E_n) \\ &= P(E_1) - \lim_{n\to\infty} P(E_n) \end{aligned} \tag{1.17}$$

Countable additivity requires disjointness of the events $E_1, E_2, \ldots \in \mathcal{E}$. In the absence of disjointness, one can still conclude *Boole's inequality*,

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} P(E_n) \tag{1.18}$$

The inequality is demonstrated by expressing the union as a disjoint union and applying countable additivity:

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = P\left(E_1 \cup \bigcup_{n=2}^{\infty} (E_1^c \cap E_2^c \cap \cdots \cap E_{n-1}^c) \cap E_n\right)$$

$$= P(E_1) + \sum_{n=2}^{\infty} P((E_1^c \cap E_2^c \cap \cdots E_{n-1}^c) \cap E_n) \tag{1.19}$$

Equation 1.18 follows because each set composing the sum is a subset of E_n.

1.1.2. Conditional Probability

Rather than simply asking the probability of an event E occurring, one might ask the probability of E occurring given that some other event F is known to have occurred. The question arises because one wishes to predict the outcome of one measurement given knowledge of one or more other measurements. If E and F are two events, then the probability of observing event E conditioned by prior knowledge that event F has occurred is defined in the following manner: if $(S, \mathcal{E}, P)$ is a probability space and $P(F) > 0$, then the *conditional probability measure* relative to F is defined by

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{1.20}$$

The definition can be motivated by the following considerations. Suppose a point is to be randomly chosen in a region R of volume $v(R) = 1$ and, for any subregion $E \subset R$, the probability of the point falling in E is given by its volume $v(E)$. If one is asked the probability of the point falling in E conditioned by the prior knowledge that it has fallen in subregion F, then it is geometrically reasonable to choose this new conditioned probability to be $v(E \cap F)/v(F)$.

Theorem 1.3. If $(S, \mathcal{E}, P)$ is a probability space and $P(F) > 0$, then $P(\cdot\,|F)$ is a probability measure on the σ-algebra $\mathcal{E}$. ■

The theorem means that $P(\cdot\,|F)$ satisfies the probability axioms. It is immediate from Eq. 1.20 that $P(E|F) \geq 0$ and, since $S \cap F = F$, that $P(S|F) = 1$. As for countable additivity, if events E_1, E_2,... are mutually disjoint, then the definition of conditional probability and the countable additivity of P yield

$$P\left(\bigcup_{n=1}^{\infty} E_n \middle| F\right) = \frac{1}{P(F)} P\left(\left(\bigcup_{n=1}^{\infty} E_n\right) \cap F\right)$$

$$= \frac{1}{P(F)} P\left(\bigcup_{n=1}^{\infty} (E_n \cap F)\right)$$

$$= \sum_{n=1}^{\infty} \frac{P(E_n \cap F)}{P(F)}$$

$$= \sum_{n=1}^{\infty} P(E_n|F) \tag{1.21}$$

Cross multiplication in Eq. 1.20 yields the *multiplication principle*:

$$P(E \cap F) = P(F)P(E|F) \tag{1.22}$$

The multiplication principle extends to $n > 2$ events: if

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) > 0 \tag{1.23}$$

then

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1, E_2) \cdots P(E_n|E_1, E_2, \ldots, E_{n-1}) \tag{1.24}$$

where $P(E_3 \mid E_1, E_2)$ denotes $P(E_3 \mid E_1 \cap E_2)$.

On many occasions one is interested in the conditional probability $P(F|E)$ but only knows $P(E|F)$. In such a situation, the following *Bayes' rule* can be applied:

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{P(E \cap F)}{P(E)} = \frac{P(F)P(E|F)}{P(E)} \tag{1.25}$$

Now suppose events $F_1, F_2, \ldots, F_n$ form a *partition* of the sample space S, meaning that the collection is disjoint and S equals the union of $F_1, F_2, \ldots, F_n$. If event $E \subset S$, then

$$P(E) = P\left(\bigcup_{k=1}^{n} (E \cap F_k)\right) = \sum_{k=1}^{n} P(E \cap F_k) \tag{1.26}$$

Putting Eqs. 1.25 and 1.26 together yields *Bayes' theorem*: if events $F_1, F_2, \ldots, F_n$ form a partition of the sample space S, event $E \subset S$, and $E, F_1, F_2, \ldots, F_n$ have positive probabilities, then

$$P(F_k|E) = \frac{P(F_k)P(E|F_k)}{\sum_{i=1}^{n} P(F_i)P(E|F_i)} \tag{1.27}$$

for $k = 1, 2,..., n$. The theorem is applied when the *prior probabilities* $P(F_i)$ and $P(E|F_i)$ composing the sum in the denominator can be obtained experimentally or from a model and we desire the *posterior probabilities* $P(F_k|E)$.

Example 1.4. A basic classification paradigm is to decide whether an object belongs to a particular class based on a measurement (or measurements) pertaining to the object. Consider observing a set of objects, say geometric shapes, and classifying an object as belonging to class C_0 or C_1 based on a real-valued measurement X. For instance X might be the perimeter, area, or number of holes of a shape. Let E be the event that $X \geq t$, where t is a fixed value, and F be the event that a randomly selected object A belongs to class C_0. Suppose we know the conditional probabilities $P(E|F)$, the probability that $X \geq t$ given $A \in C_0$, and $P(E|F^c)$, the probability that $X \geq t$ given $A \in C_1$, and we also know the probability $P(F)$ that a randomly selected object belongs to C_0. For deciding whether or not a selected object belongs to C_0, we would like to know $P(F|E)$, the probability $A \in C_0$ given $X \geq t$. $P(F|E)$ is given by Bayes' theorem:

$$P(F|E) = \frac{P(F)P(E|F)}{P(F)P(E|F) + P(F^c)P(E|F^c)}$$ ■

Events E and F are said to be *independent* if

$$P(E \cap F) = P(E)P(F) \tag{1.28}$$

Otherwise they are *dependent*. If $P(F) > 0$, then E and F are independent if and only if $P(E|F) = P(E)$. If E and F are independent, then so too are E and F^c, E^c and F, and E^c and F^c. More generally, events $E_1, E_2,..., E_n$ are independent if, for any subclass $\{E_{i_1}, E_{i_2}, \ldots, E_{i_m}\} \subset \{E_1, E_2,..., E_n\}$,

$$P\left(\bigcap_{j=1}^{m} E_{i_j}\right) = \prod_{j=1}^{m} P(E_{i_j}) \tag{1.29}$$

Note that pairwise independence of $E_1, E_2, \ldots, E_n$, namely, that each pair within the class satisfies Eq. 1.28, does not imply independence of the full class.

Example 1.5. Suppose m components $C_1, C_2, \ldots, C_m$ compose a system, F_k is the event that component C_k fails during some stated period of operation, F is the event that the system fails during the operational period, and component failures are independent. The components are said to be arranged in *series* if the system fails if any component fails and to be arranged in *parallel* if the system fails if and only if all components fail. If the system is arranged in series, then

$$F = \bigcup_{k=1}^{m} F_k$$

$$P(F) = 1 - P(F^c) = 1 - P\left(\bigcap_{k=1}^{m} F_k^c\right) = 1 - \prod_{k=1}^{m} (1 - P(F_k))$$

If the series is arranged in parallel, then

$$P(F) = P\left(\bigcap_{k=1}^{m} F_k\right) = \prod_{k=1}^{m} P(F_k) \qquad \blacksquare$$

1.2. Random Variables

Measurement randomness results from both the inherent randomness of phenomena and variability within observation and measurement systems. Quantitative description of random measurements is embodied in the concept of a random variable. The theory of random processes concerns random variables defined at points in time or space, as well as interaction between random variables.

1.2.1. Probability Distributions

Given a probability space $(S, \mathcal{E}, P)$, a *random variable* is a mapping X: $S \to \Re$, the space of real numbers, such that

$$X^{-1}((-\infty, x]) = \{z \in S: X(z) \leq x\} \tag{1.30}$$

is an element of $\mathcal{E}$ (an event) for any $x \in \Re$. If $X^{-1}((-\infty, x]) \in \mathcal{E}$ for any $x \in \Re$ (if X is a random variable), then it can be shown that $X^{-1}(B) \in \mathcal{E}$ for any Borel set $B \subset \Re$, which means in particular that $X^{-1}(B) \in \mathcal{E}$ if B is an open set, a closed set, an intersection of open sets, or a union of closed sets.

Theorem 1.4. A random variable X on a probability space $(S, \mathcal{E}, P)$ induces a probability measure P_X on the Borel σ-algebra $\mathcal{B}$ in $\mathfrak{R}$ by

$$P_X(B) = P(X^{-1}(B)) = P(\{z \in S: X(z) \in B\}) \quad \blacksquare \tag{1.31}$$

To prove the theorem, first note that $P_X(B)$ is defined for any $B \in \mathcal{B}$ because $X^{-1}(B) \in \mathcal{E}$ for any $B \in \mathcal{B}$. The first two probability axioms are easily verified:

$$P_X(\mathfrak{R}) = P(X^{-1}(\mathfrak{R})) = P(S) = 1 \tag{1.32}$$

$$P_X(B^c) = P(X^{-1}(B^c)) = P([X^{-1}(B)]^c) = 1 - P(X^{-1}(B)) = 1 - P_X(B) \tag{1.33}$$

If $B_1, B_2, \ldots$ form a disjoint countable collection of Borel sets, then $X^{-1}(B_1)$, $X^{-1}(B_2), \ldots$ form a disjoint countable collection of events in $\mathcal{E}$. Hence,

$$\begin{aligned} P_X\left(\bigcup_{i=1}^{\infty} B_i\right) &= P\left(X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right)\right) \\ &= P\left(\bigcup_{i=1}^{\infty} X^{-1}(B_i)\right) \\ &= \sum_{i=1}^{\infty} P(X^{-1}(B_i)) \\ &= \sum_{i=1}^{\infty} P_X(B_i) \end{aligned} \tag{1.34}$$

Hence, the three probability axioms are satisfied by the induced probability measure.

The induced probability measure of a random variable X defines the inclusion probability $P(X \in B)$ by

$$P(X \in B) = P_X(B) = P(X^{-1}(B)) \tag{1.35}$$

***Example* 1.6.** Let $(S, \mathcal{E}, P)$ be a probability space, E_0 and E_1 partition S, and random variable X be defined by $X(a) = 0$ if $a \in E_0$ and $X(a) = 1$ if $a \in E_1$. For any Borel set B, $P_X(B) = 0$, if $\{0, 1\} \cap B = \varnothing$, $P_X(B) = P(E_0)$, if $\{0, 1\} \cap B = \{0\}$, $P_X(B) = P(E_1)$, if $\{0, 1\} \cap B = \{1\}$, $P_X(B) = 1$, if $\{0, 1\} \cap B = \{0, 1\}$. $\blacksquare$

As a consequence of its inducing a probability measure on the Borel σ-algebra, the random variable X induces a probability space $(\mathfrak{R}, \mathcal{B}, P_X)$ on the real line. If we are only concerned with X and its inclusion probabilities $P(X \in B)$ for Borel sets, once we have a representation for P_X we need not concern ourselves with the original sample space. This observation is the key to probabilistic modeling: when measuring random phenomena we need only model the distribution of probability mass over the real line associated with the random variable.

For a random variable X defined on the probability space $(S, \mathcal{E}, P)$, define its *probability distribution function* F_X: $\mathfrak{R} \to [0, 1]$ by

$$F_X(x) = P(X \le x) = P_X((-\infty, x]) \tag{1.36}$$

Interval probabilities are expressed via the probability distribution function. If $a < b$, then

$$\begin{aligned} P(a < X \le b) &= F_X(b) - F_X(a) \\ P(a \le X \le b) &= F_X(b) - F_X(a) + P(X = a) \\ P(a < X < b) &= F_X(b) - F_X(a) - P(X = b) \\ P(a \le X < b) &= F_X(b) - F_X(a) + P(X = a) - P(X = b) \end{aligned} \tag{1.37}$$

Theorem 1.5. If F_X is the probability distribution function for a random variable X, then

(i) F_X is increasing.
(ii) F_X is continuous from the right.
(iii) $\lim_{x \to -\infty} F_X(x) = 0$.
(iv) $\lim_{x \to \infty} F_X(x) = 1$.

Conversely, if F is any function satisfying the four properties, then there exists a probability space and a random variable X on that space such that the probability distribution function for X is given by F. ■

We demonstrate that the four conditions stated in Theorem 1.5 hold whenever F_X is a probability distribution function. First, if $x_1 \le x_2$, then $(-\infty, x_1] \subset (-\infty, x_2]$, implying that

$$F_X(x_1) = P_X((-\infty, x_1]) \le P_X((-\infty, x_2]) = F_X(x_2) \tag{1.38}$$

so that F_X is increasing. To show continuity from the right, suppose $\{x_n\}$ is a decreasing sequence and $x_n \to x$. Then, owing to continuity from above,

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P_X((-\infty, x_n])$$

$$= P_X\left(\bigcap_{n=1}^{\infty}(-\infty, x_n]\right) \tag{1.39}$$

$$= P_X((-\infty, x])$$

which is $F_X(x)$, thereby demonstrating continuity from the right. For property (iii), suppose $\{x_n\}$ is a decreasing sequence with $x_n \to -\infty$. Then, as in Eq. 1.39,

$$\lim_{n\to\infty} F_X(x_n) = P_X\left(\bigcap_{n=1}^{\infty}(-\infty, x_n]\right) = P_X(\varnothing) = 0 \tag{1.40}$$

For property (iv), suppose $\{x_n\}$ is an increasing sequence with $x_n \to \infty$. Then, owing to continuity from below,

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P_X((-\infty, x_n])$$

$$= P_X\left(\bigcup_{n=1}^{\infty}(-\infty, x_n]\right) \tag{1.41}$$

$$= P_X(\Re)$$

which is 1. We will not prove the converse of the theorem because it requires the theory of Lebesgue-Stieltjes measure.

Owing to the converse of Theorem 1.5, once a probability distribution function is defined, *ipso facto* there exists a random variable whose behavior is described by the function. The probability distribution function is then said to be the *law* of the random variable and it is chosen to model some phenomena. Unless there is potential confusion, we will drop the subscript when denoting a probability distribution function. The terminology *probability distribution* is applied to either the probability distribution function or to a random variable having the probability distribution function. Two random variables possessing the same probability distribution function are said to be *identically distributed* and are probabilistically indistinguishable.

1.2.2. Probability Densities

A probability distribution is commonly specified by a nonnegative function $f(x)$ for which

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{1.42}$$

Such a function is called a *probability density* and yields a probability distribution function (and therefore a corresponding random variable) by

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \tag{1.43}$$

$F(x)$ is a continuous function and models the probability mass of a random variable X taking values in a continuous range. $F(x)$, $f(x)$, and X are said to constitute a *continuous probability distribution*. According to the fundamental theorem of calculus,

$$F'(x) = \frac{d}{dx}F(x) = f(x) \tag{1.44}$$

at any point x where the density f is continuous.

According to Eq. 1.37, for $a < b$,

$$P(X = b) = F(b) - F(a) - P(a < X < b) \tag{1.45}$$

Owing to continuity from above, $P(a < X < b) \to 0$ as $a \to b$ from the left. Because F is continuous, $F(b) - F(a) \to 0$ as $a \to b$ from the left. Hence, $P(X = b) = 0$. Since the point b was chosen arbitrarily, we conclude that all point probabilities of a continuous probability distribution are zero. Consequently, for a continuous distribution, all interval-probability expressions of Eq. 1.37 reduce to the first, which now becomes

$$P(a < X \le b) = \int_{a}^{b} f(t)\,dt \tag{1.46}$$

Example 1.7. For $a < b$, the *uniform distribution* over interval $[a, b]$ is characterized by the probability density $f(x) = 1/(b - a)$ for $a \le x \le b$ and $f(x) = 0$ otherwise. The corresponding probability distribution function is defined by $f(x) = 0$ for $x < a$,

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

for $a \le x \le b$, and $f(x) = 1$ for $x > b$. As determined by the density, the mass of the distribution is spread uniformly over $[a, b]$ and the probability distribution function ramps from 0 to 1 across $[a, b]$. ■

A discrete random variable X is modeled by specifying a countable *range* of points $\Omega_X = \{x_1, x_2, \ldots\}$ and a nonnegative *probability mass function* (*discrete density*) $f(x)$ such that $f(x) > 0$ if and only if $x \in \Omega_X$ and

$$\sum_{k=1}^{\infty} f(x_k) = 1 \tag{1.47}$$

The probability mass function generates a probability distribution function by

$$F(x) = \sum_{\{k:x_k \le x\}} f(x_k) \tag{1.48}$$

$F(x)$ is a step function with jump $f(x_k)$ at x_k and, assuming $x_1 < x_2 < \ldots$, $F(x)$ is constant on the interval $[x_k, x_{k+1})$. A random variable X with probability distribution function $F(x)$ has point probabilities $P(X = x_k) = f(x_k)$. The first interval probability of Eq. 1.37 becomes

$$P(a < X \le b) = \sum_{\{k:a<x_k \le b\}} f(x_k) \tag{1.49}$$

Other interval probabilities differ depending on whether $f(a) > 0$ or $f(b) > 0$. If the point set Ω_X is finite, then the sum of Eq. 1.48 is finite and $F(x) = 1$ for sufficiently large x. A probability mass function and its random variable constitute a *discrete distribution*.

To unify the representation of continuous and discrete distributions, we employ delta functions to represent discrete distributions. As a generalized function, a delta function is an operator on functions; nevertheless, we take a view often adopted in engineering and treat a delta function $\delta(x)$ as though it were a function with the integration property

$$\int_{-\infty}^{\infty} g(x)\delta(x-a)\,dx = g(a) \tag{1.50}$$

Under this convention we apply Eq. 1.44 in a generalized sense, and the probability mass function $f(x)$ corresponding to the point set Ω_X can be represented as

$$f(x) = \sum_{k=1}^{\infty} f(x_k)\delta(x - x_k) \tag{1.51}$$

Use of generalized functions for discrete densities is theoretically justified. For applications, one need only recognize that use of delta functions according to Eq. 1.51 is appropriate so long as Ω_X does not have any limit points.

There exist probability distributions that are neither continuous nor discrete. These *mixed* distributions have contributions from both continuous and discrete parts. Although we will not explicitly employ them as models in the text, they constitute valid distributions and need to be considered in the general theory. The salient point is that Theorem 1.5 determines the central role of probability distribution functions and their properties. We will take a dual approach in this regard. We will utilize Eq. 1.44 in a generalized sense, so that the theory is unified in the framework of probability densities, but we will often prove theorems assuming a continuous distribution, thereby avoiding theoretical questions concerning generalized functions. A more direct and unified approach would be to remain in the context of probability distribution functions; however, this would require the use of Lebesgue-Stieltjes integration and force us into measure-theoretic questions outside the scope of the text. For those with a background in measure theory, recall that according to the Lebesgue decomposition of a function F satisfying the conditions of Theorem 1.5, F is differentiable almost everywhere and F is decomposed as $F = F_s + F_a$, where F_a is absolutely continuous, F_s is singular [$F_s' = 0$ almost everywhere], and F_s can be continuous and not be a constant.

1.2.3. Functions of a Random Variable

We are rarely concerned with a random variable in isolation, but rather with its relations to other random variables. The terminology "variable" applies because we usually consider a system with one or more random inputs and one or more random outputs. The simplest case involves a function of a single random variable.

For a discrete random variable X, the density of a function $Y = g(X)$ is given by

$$f_Y(y) = P(Y = y) = \sum_{\{x:g(x)=y\}} f_X(x) \tag{1.52}$$

In particular, if g is one-to-one, then $\{x: g(x) = y\}$ consists of the single element $g^{-1}(y)$ and

$$f_Y(y) = f_X[g^{-1}(y)] \tag{1.53}$$

Finding output distributions of continuous random variables is more difficult. One approach is to find the probability distribution function of the output random variable and then differentiate to obtain the output density.

Example 1.8. Consider the affine transformation $Y = aX + b$, $a \neq 0$. For $a > 0$,

$$F_Y(y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

For $a < 0$,

$$F_Y(y) = P\left(X \geq \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right)$$

Differentiation for $a > 0$ yields

$$f_Y(y) = \frac{d}{dx} F_X\left(\frac{y-b}{a}\right) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$$

Except for a minus sign in front, the same expression is obtained for $a < 0$. Combining the two results yields

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \qquad \blacksquare$$

Suppose $y = g(x)$ is differentiable for all x and has a derivative that is either strictly greater than or strictly less than 0. Then $x = g^{-1}(y)$ has a derivative known as the *Jacobian* of the transformation and denoted by $J(x; y)$. Moreover, there exist values y_1 and y_2, $y_1 < y_2$, either or both possibly infinite, such that for any y between y_1 and y_2, there exists exactly one value x such that $y = g(x)$.

Theorem 1.6. If X is a continuous random variable and $y = g(x)$ has a derivative that is either strictly greater than or strictly less than 0, then $Y = g(X)$ is a continuous random variable with density

$$f_Y(y) = \begin{cases} f_X[g^{-1}(y)]|J(x; y)|, & \text{if } y_1 < y < y_2 \\ 0, & \text{otherwise} \end{cases} \tag{1.54}$$

where y_1 and y_2 are the limiting values described prior to the theorem. $\blacksquare$

To prove the theorem, suppose $y_1 < y < y_2$ and $g' > 0$. Then

$$F_Y(y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y)) \tag{1.55}$$

Differentiation with respect to y yields

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y) = f_X(g^{-1}(y))J(x;y) \tag{1.56}$$

Now suppose $g' < 0$. Then

$$F_Y(y) = P(g(X) \le y) = P(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y)) \tag{1.57}$$

Differentiation yields

$$f_Y(y) = -f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y) = f_X(g^{-1}(y))[-J(x;y)] \tag{1.58}$$

Since $g' < 0$, $J(x;y) < 0$, so that $-J(x;y) = |J(x;y)|$. Combining Eqs. 1.56 and 1.58 yields the result for $y_1 < y < y_2$. It is straightforward to show that $f_Y(y) = 0$ for $y \notin (y_1, y_2)$.

***Example* 1.9.** The exponential function $y = g(x) = e^{tx}$, $t > 0$, satisfies the preceding conditions. For $y > 0$,

$$g^{-1}(y) = \frac{\log y}{t}$$

$$J(x;y) = \frac{d}{dt}\left(\frac{\log y}{t}\right) = \frac{1}{ty}$$

$$f_Y(y) = \frac{1}{ty} f_X\left(\frac{\log y}{t}\right)$$

For $y < 0$, $f_Y(y) = 0$. For $y = 0$, $f_Y(0)$ can be defined arbitrarily. ■

1.3. Moments

Full description of a random variable requires characterization of its probability distribution function; however, most applications involve only partial description of a random variable via its moments. We state definitions and properties in terms of densities, for which moments have geometric

intuition relative to probability mass. Definitions involving integrals can be interpreted directly for continuous random variables; for discrete random variables they can be interpreted using delta functions or can be re-expressed as sums.

1.3.1. Expectation and Variance

The *expected value* (*expectation*) of a random variable X with density $f(x)$ is defined by

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx \tag{1.59}$$

as long as the defining integral is absolutely convergent, meaning

$$\int_{-\infty}^{\infty} |x| f(x)\, dx < \infty \tag{1.60}$$

Whenever we write $E[X]$, it is implicitly assumed that the defining integral is absolutely convergent. The expected value is the center of mass for the probability distribution (density). It is also called the *mean* and denoted by μ_X (or simply μ if X is clear from the context).

To apply Eq. 1.59 directly to a function $g(X)$ of a random variable would require first finding the density of $g(X)$. Fortunately, this need not be done.

Theorem 1.7. If $g(x)$ is any piecewise continuous, real-valued function and X is a random variable possessing density $f(x)$, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \qquad \blacksquare \tag{1.61}$$

We demonstrate the theorem for the special case in which g is differentiable and has a derivative that is either strictly greater than or strictly less than 0. By Theorem 1.6,

$$E[g(X)] = \int_{y_1}^{y_2} y f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y)\, dy \tag{1.62}$$

For the substitution $y = g(x)$, $dy = g'(x)dx$ and $x = g^{-1}(y)$. According to the definition of y_1 and y_2, as y varies from y_1 to y_2, x goes from $-\infty$ to ∞. Thus, the substitution yields

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \frac{d}{dy} g^{-1}(y)\, g'(x) dx \tag{1.63}$$

Equation 1.61 follows because

$$\frac{d}{dy} g^{-1}(y) = \frac{1}{g'(x)} \tag{1.64}$$

Assuming the defining integral is absolutely convergent, for integer $k \geq 1$, the kth *moment* about the origin of a random variable X possessing density $f(x)$ is defined by

$$\mu_k' = E[X^k] = \int_{-\infty}^{\infty} x^k f(x)\, dx \tag{1.65}$$

$E[X]$ is the first moment of X. Assuming the defining integral is absolutely convergent, the kth *central moment* is defined by

$$\mu_k = E[(X - \mu)^k] = \int_{-\infty}^{\infty} (x - \mu)^k f(x)\, dx \tag{1.66}$$

where μ is the mean of X. Note that Theorem 1.7 was invoked in defining both moments and central moments. The kth central moment is the kth moment of the *centered* random variable $X - \mu$, whose mean is 0.

The second central moment is called the *variance* and is defined by

$$\sigma^2 = \mu_2 = E[(X - \mu)^2] \tag{1.67}$$

The variance measures the spread of the probability mass about the mean: the more dispersed the mass, the greater the variance. To specify the random variable X when writing the variance, we write $\mathrm{Var}[X]$ or σ_X^2. The square root of the variance is called the *standard deviation*.

Rather than compute the variance by its defining integral, it is usually more easily computed via the relation

$$\sigma^2 = \mu_2' - \mu^2 \tag{1.68}$$

which is obtained by expanding the integral defining $E[(X - \mu)^2]$. It is straightforward to show that, for any constants a and b,

$$\text{Var}[aX + b] = a^2\text{Var}[X] \tag{1.69}$$

***Example* 1.10.** Consider the uniform distribution over the interval $[a, b]$ defined in Example 1.7. For $k = 1, 2,...,$ the kth moment is

$$\mu_k' = E[X^k] = \frac{1}{b-a}\int_a^b x^k\,dx = \frac{b^{k+1} - a^{k+1}}{(k+1)(b-a)}$$

The mean and second moment are found by substituting $k = 1$ and $k = 2$, respectively, into μ_k': $\mu = (a + b)/2$, $\mu_2' = (b^2 + ab + a^2)/3$. Equation 1.68 yields $\sigma^2 = (b - a)^2/12$. ■

The next theorem provides two inequalities bounding the probability mass of a random variable over a range of values. The first (the *generalized Chebyshev inequality*) bounds the probability mass of a nonnegative random variable above a given value in terms of its expectation; the second (*Chebyshev's inequality*) quantifies the manner in which the variance measures the absolute deviation of a random variable from its mean. Chebyshev's inequality plays important roles in statistical estimation and the convergence of sequences of random variables.

Theorem 1.8. If random variable X is nonnegative and has mean μ, then, for any $t > 0$,

$$P(X \geq t) \leq \frac{\mu}{t} \tag{1.70}$$

If X (not necessarily nonnegative) possesses mean μ and variance σ^2, then, for any $t > 0$,

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \blacksquare \tag{1.71}$$

For a continuous random variable, the generalized Chebyshev inequality results from the following inequality upon division by t:

$$\mu = \int_0^\infty xf(x)\,dx$$

$$\geq \int_t^\infty xf(x)\,dx$$

$$\geq t\int_t^\infty f(x)\,dx \tag{1.72}$$

$$= tP(X \geq t)$$

Applying the generalized Chebyshev inequality to $|X - \mu|^2$ and t^2 yields the ordinary form in the following way:

$$P(|X-\mu| \geq t) = P(|X-\mu|^2 \geq t^2) \leq \frac{E[|X-\mu|^2]}{t^2} = \frac{\sigma^2}{t^2} \tag{1.73}$$

Letting $t = k\sigma$, $k > 0$, in Eq. 1.71 expresses Chebyshev's inequality in terms of the number of standard deviations from the mean:

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{1.74}$$

A different form results from complementation in Eq. 1.71,

$$P(|X-\mu| < t) \geq 1 - \frac{\sigma^2}{t^2} \tag{1.75}$$

Thus, the probability mass over the interval $(\mu - t, \mu + t)$ is bounded from below. For small variance, the mass is tightly concentrated about the mean. The Chebyshev inequality does not take into account the actual distribution and therefore it is often rather loose; however, without a strengthened hypothesis it cannot be improved.

1.3.2. Moment-Generating Function

Using an exponential transform can help with some tasks involving probability densities. The moment-generating function of a random variable X having density $f(x)$ is defined by

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \tag{1.76}$$

for all t for which the integral is finite. For any constants a and b, straightforward use of the properties of the exponential shows that

$$M_{aX+b}(t) = e^{bt} M_X(at) \tag{1.77}$$

To be useful, a transform requires unique inversion. The next theorem is the *uniqueness* theorem for the moment-generating function.

Theorem 1.9. If $M_X(t) = M_Y(t)$ for all t in some open interval containing $t = 0$, then the random variables X and Y are identically distributed. ■

The moment-generating function can be used to find moments. Suppose X has density $f(x)$ and the kth moment of X exists. Taking the derivative of $M_X(t)$ with respect to t yields

$$\begin{aligned} \frac{d}{dt} M_X(t) &= \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \\ &= \int_{-\infty}^{\infty} \frac{\partial}{\partial t} [e^{tx} f(x)]\, dx \\ &= \int_{-\infty}^{\infty} x e^{tx} f(x)\, dx \end{aligned} \tag{1.78}$$

where we have assumed in the second equality that the derivative can be brought inside the integral (which it can be for all distributions in which the moment-generating function will be invoked in this text). Proceeding by recursive differentiation yields

$$M_X^{(k)}(t) = \int_{-\infty}^{\infty} x^k e^{tx} f(x)\, dx \tag{1.79}$$

Letting $t = 0$ gives

$$M_X^{(k)}(0) = \int_{-\infty}^{\infty} x^k f(x)\,dx = \mu_k' \tag{1.80}$$

where $M_X^{(k)}(0)$ is the kth derivative of $M_X(t)$ with respect to t, evaluated at $t = 0$.

Example 1.11. For $b > 0$, the *exponential distribution* is characterized by the density $f(x) = be^{-bx}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$, where $b > 0$. Its moment-generating function is

$$M_X(t) = \int_0^{\infty} e^{tx} be^{-bx}\,dx = b\int_0^{\infty} e^{-(b-t)x}\,dx = \frac{b}{b-t}$$

for $t < b$ (so that the integral is finite). Successive differentiations with respect to t yield

$$M_X^{(k)}(t) = \frac{k!b}{(b-t)^{k+1}}$$

for $k = 1, 2,\ldots$. Letting $t = 0$ yields $\mu_k' = k!/b^k$. Hence, the mean, second moment, and variance are $\mu = 1/b$, $\mu_2' = 2/b^2$, and $\sigma^2 = 1/b^2$, respectively. ■

1.4. Important Probability Distributions

This section provides definitions and properties of some commonly employed probability distributions. The binomial distribution describes probabilities associated with repeated, independent binary trials; the Poisson distribution models a fundamental class of random point processes; the normal distribution is used extensively in statistics, serves as a model for noise in image and signal filtering, and is a limiting distribution in many key circumstances; the gamma distribution governs a family of many useful distributions and is useful for modeling grain sizing and interarrival times for queues; the beta distribution takes on many types of shapes for different combinations of its parameters and is therefore useful for modeling various kinds of phenomena.

1.4.1. Binomial Distribution

Suppose an experiment consists of n trials, $n > 0$. The trials are called *Bernoulli trials* if three conditions are satisfied: (1) each trial is defined by a sample space $\{s, f\}$ having two outcomes, called *success* and *failure*; (2) there is a number p, $0 < p < 1$, such that, for each trial, $P(s) = p$ and $P(f) = q = 1 - p$;

and (3) the trials are independent. Bernoulli trials are realized by selecting n balls randomly, in succession, and with replacement after each selection from an urn containing k black balls and m white balls. If selecting a black ball constitutes a success, then $p = k/(k + m)$. The appropriate sample space for an experiment composed of n Bernoulli trials is $\{s, f\}^n$, the Cartesian product of $\{s, f\}$ with itself n times. Let the random variable X count the number of successes in the n trials. Its probability mass function is nonzero for $x = 0, 1,..., n$. The probability of any outcome $(u_1, u_2,..., u_n) \in \{s, f\}^n$ for which exactly x of the components equal s is $p^x q^{n-x}$. There are

$$C_{n,x} = \binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{1.81}$$

such outcomes. Hence, the probability mass function for X is

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x} \tag{1.82}$$

for $x = 0, 1,..., n$. Any random variable having this density is said to possess a *binomial distribution* and is said to *binomially distributed.* The binomial probability distribution function is

$$F(x) = \sum_{k \le x} \binom{n}{k} p^k q^{n-k} \tag{1.83}$$

$F(x) = 0$ for $x < 0$, $F(x) = 1$ for $x \ge n$, $F(x)$ has jumps at $x = 0, 1,..., n$, and $F(x)$ is constant on all intervals $[x, x + 1)$ for $x = 0, 1,..., n - 1$.

The moment-generating function for the binomial distribution is given by

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x} \\ &= (pe^t + q)^n \end{aligned} \tag{1.84}$$

where the last equality follows from the recognition that the preceding sum is the binomial expansion of the last expression. Taking the first and second derivative of the moment-generating function with respect to t and then

setting $t = 0$ yields the mean and second moment according to Eq. 1.80, and application of Eq. 1.68 then gives the variance: $\mu = np$, $\mu_2' = np(1 + np - p)$, $\sigma^2 = npq$.

1.4.2. Poisson Distribution

Whereas the binomial distribution can be arrived at by means of an experimental protocol involving a finite sample space, the Poisson distribution cannot be so developed. Later in the text we will show how the Poisson distribution results from an arrival process in time; for now we define it and give some basic properties. A discrete random variable X is said to possess a *Poisson distribution* if it has probability mass function

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{1.85}$$

for $x = 0, 1, 2,...$, where $\lambda > 0$. The Poisson probability distribution function is

$$F(x) = \sum_{k \le x} \frac{e^{-\lambda}\lambda^k}{k!} \tag{1.86}$$

$F(x) = 0$ for $x < 0$ and $F(x)$ has jumps at $x = 0, 1, 2,....$

The moment-generating function for the Poisson distribution is given by

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{\infty} \frac{e^{tx}e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= \exp[\lambda(e^t - 1)] \end{aligned} \tag{1.87}$$

the last equality following from the fact that the series is the Taylor series for the exponential function. The mean, second moment, and variance are found from the moment-generating function: $\mu = \lambda$, $\mu_2' = \lambda + \lambda^2$, $\sigma^2 = \lambda$.

The binomial and Poisson distributions are asymptotically related. Let $b(x; n, p)$ and $\pi(x; \lambda)$ denote the binomial and Poisson densities, respectively. Then, for $x = 0, 1, 2,...$,

$$\lim_{n\to\infty} b\left(x; n, \frac{\lambda}{n}\right) = \pi(x; \lambda) \tag{1.88}$$

This asymptotic relation can be used to approximate the binomial distribution by the Poisson distribution. Replacing λ/n by p in Eq. 1.88 yields

$$\lim_{n\to\infty} b(x; n, p) = \pi(x; np) \tag{1.89}$$

so that for large n, $b(x; n, p) \approx \pi(x; np)$. There is a caveat in applying this approximation: p must be sufficiently small so that np is not too large, even though n is large. The reason for this caveat has to do with the rate of convergence to the limit.

To obtain the limit in Eq. 1.88, first note that

$$\lim_{n\to\infty} \frac{n(n-1)\cdots(n-x+1)}{n^x} = \prod_{k=0}^{x-1} \lim_{n\to\infty}\left(1-\frac{k}{n}\right) = 1 \tag{1.90}$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} = 1 \tag{1.91}$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n} = e^{-\lambda} \tag{1.92}$$

Writing out $b(x; n, \lambda/n)$, rearranging terms, and taking the product of the limits yields the desired limit relation:

$$\begin{aligned}
\lim_{n\to\infty} b\left(x; n, \frac{\lambda}{n}\right) &= \lim_{n\to\infty}\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n\to\infty}\frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n\to\infty}\frac{n(n-1)\cdots(n-x+1)}{n^x}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x e^{-\lambda}}{x!}
\end{aligned} \tag{1.93}$$

***Example* 1.12.** Suppose that for a particular communication channel the error rate is 1 incorrect transmission per 100 messages and transmissions are independent. If n messages are sent, then, in essence, there are n Bernoulli trials and the probability of a success, which is actually an erroneous

transmission, is $p = 0.01$. Letting X denote the number of erroneous transmissions in the n messages, the Poisson approximation is

$$P(X=x) \approx \frac{e^{-np}(np)^x}{x!}$$

For $n = 200$, $P(X \geq 3) = 0.3233$ by both methods. ■

1.4.3. Normal Distribution

A continuous random variable X is said to possess a *normal* (*Gaussian*) *distribution* if it has the density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1.94}$$

where $-\infty < x < \infty$, $-\infty < \mu < \infty$, and $\sigma > 0$. Its probability distribution function is

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \tag{1.95}$$

In the parameterization of Eq. 1.94, the mean and variance of the normal distribution are μ and σ^2, respectively, as we will soon show.

Setting $\mu = 0$ and $\sigma = 1$ yields the *standard normal distribution*, whose density and probability distribution function are given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{1.96}$$

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} dy \tag{1.97}$$

respectively. $\phi(z)$ is a legitimate density because

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = 1 \tag{1.98}$$

The integral from $-\infty$ to ∞ of the normal density of Eq. 1.94 can be shown to equal 1 by reducing it to the integral of Eq. 1.98 via the substitution $z =$

$(x - \mu)/\sigma$. In fact, the transformation $Z = (X - \mu)/\sigma$ transforms a normally distributed random variable X into a standard normal random variable. This is demonstrated by applying the results of Example 1.8 to the normal density of Eq. 1.94 with $a = 1/\sigma$ and $b = -\mu/\sigma$. Moreover, for any c and d,

$$P(c < X < d) = \Phi\left(\frac{d-\mu}{\sigma}\right) - \Phi\left(\frac{c-\mu}{\sigma}\right) \tag{1.99}$$

Hence, all interval probabilities for a normally distributed random variable can be found from the standard-normal probability distribution function, whose values can be found in standard-normal statistical tables.

The moment-generating function of the normal distribution is derived in the following manner:

$$\begin{aligned} M_X(t) &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\,dx \\ &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{x^2 - 2\mu x + \mu^2 - 2\sigma^2 tx}{\sigma^2}\right)\right]dx \\ &= \exp\left[\mu t + \frac{t^2\sigma^2}{2}\right]\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{x-(\mu+t\sigma^2)}{\sigma}\right)^2\right]dx \\ &= \exp\left[\mu t + \frac{t^2\sigma^2}{2}\right]\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy \\ &= \exp\left[\mu t + \frac{t^2\sigma^2}{2}\right] \end{aligned} \tag{1.100}$$

where the third equality results from completing the square in the exponential, the fourth results from the substitution $y = [x - (\mu + t\sigma^2)]/\sigma$, and the last from Eq. 1.98. Obtaining the derivatives of the moment-generating function at $t = 0$ and applying Eqs. 1.68 and 1.80 shows that the mean and variance of the normal distribution are μ and σ^2, respectively.

1.4.4. Gamma Distribution

The gamma distribution involves the *gamma function*, which is defined for $x > 0$ by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt \tag{1.101}$$

For $x > 0$, $\Gamma(x + 1) = x\Gamma(x)$. Thus, the gamma function is a generalization of the factorial function. In fact, if x is an integer, then $\Gamma(x + 1) = x!$. A random variable X is said to possess a *gamma distribution* with parameters $\alpha > 0$ and $\beta > 0$ if it has the density

$$f(x) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \tag{1.102}$$

for $x \geq 0$ and $f(x) = 0$ for $x < 0$. Its probability distribution function is given by $F(x) = 0$ for $x \leq 0$ and

$$F(x) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-t/\beta}\, dt \tag{1.103}$$

for $x > 0$. Some gamma densities are shown in Fig. 1.1.

For $t < 1/\beta$, the moment-generating function is given by

$$\begin{aligned} M_X(t) &= \frac{\beta^{-\alpha}}{\Gamma(\alpha)} \int_{-\infty}^\infty x^{\alpha-1} e^{-[(1/\beta)-t]x}\, dx \\ &= \frac{\beta^{-\alpha}}{\Gamma(\alpha)} \left(\frac{1}{\beta} - t\right)^{-\alpha} \int_{-\infty}^\infty u^{\alpha-1} e^{-u}\, du \\ &= (1 - \beta t)^{-\alpha} \end{aligned} \tag{1.104}$$

where the first integral is finite for $t < 1/\beta$, the second follows from the substitution $u = [(1/\beta) - t]x$, and the final equality follows from the definition of $\Gamma(\alpha)$.

Differentiating the moment-generating function k times in succession yields

$$M_X^{(k)}(t) = \beta^k(\alpha + k - 1)\cdots(\alpha + 1)\alpha(1 - \beta t)^{-\alpha-k} \tag{1.105}$$

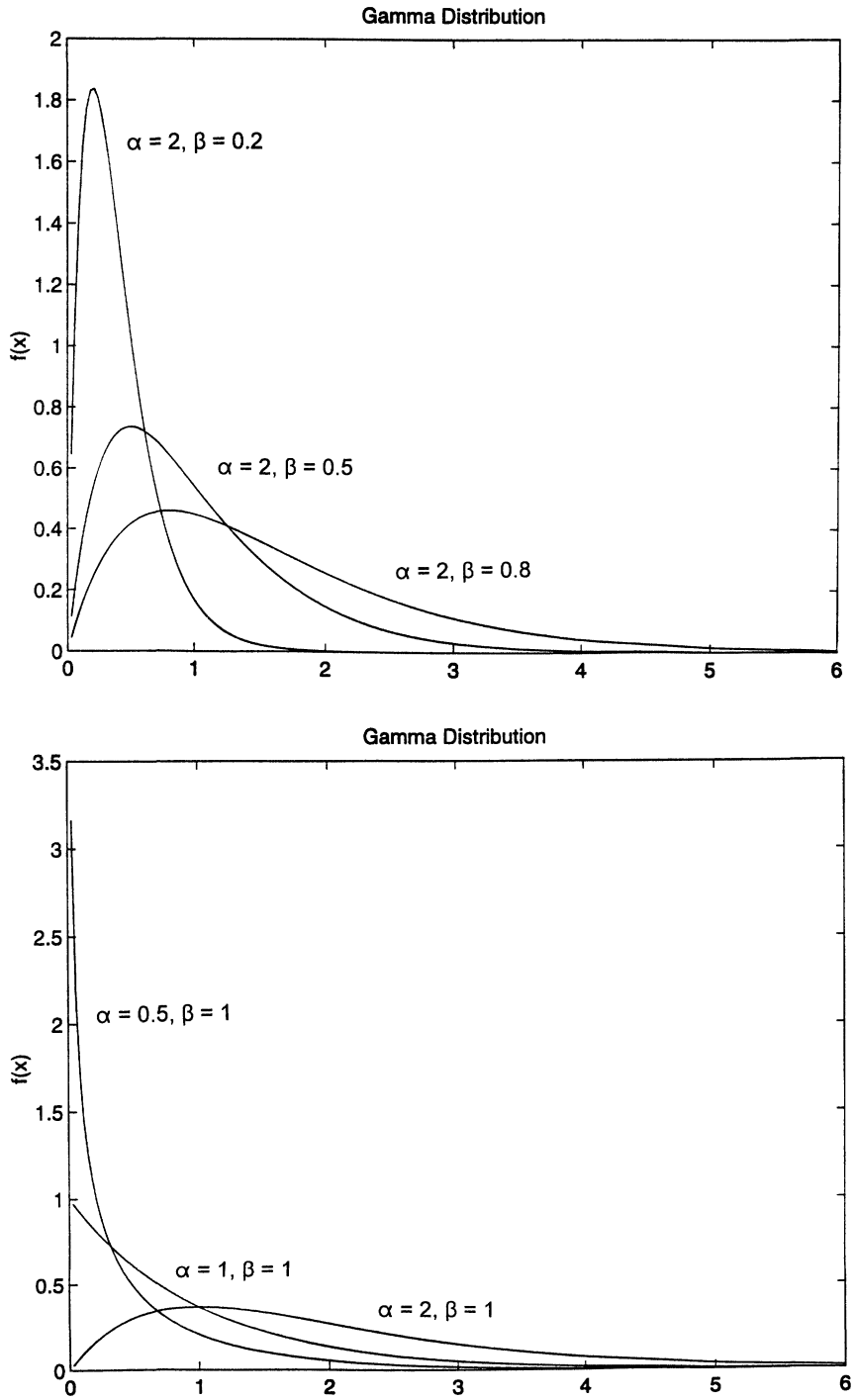

Figure 1.1 Gamma densities.

Letting $t = 0$ yields

$$\mu_k' = \beta^k(\alpha + k - 1)\cdots(\alpha + 1)\alpha = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)}\beta^k \tag{1.106}$$

Hence, $\mu = \alpha\beta$, $\mu_2' = \beta^2(\alpha + 1)\alpha$, $\sigma^2 = \alpha\beta^2$.

When $\alpha = k$ is an integer, the gamma distribution is sometimes called a *k-Erlang distribution* and it then takes the form

$$f(x) = \frac{\beta^{-k}}{(k-1)!}x^{k-1}e^{-x/\beta} \tag{1.107}$$

for $x \geq 0$ and $f(x) = 0$ for $x < 0$.

The exponential density, introduced in Example 1.11, results from the gamma density by letting $\alpha = 1$ and $\beta = 1/b$. As shown in the example, the exponential density has mean $\mu = 1/b$ and variance $\sigma^2 = 1/b^2$. A salient property of the exponential distribution is that it is memoryless: a random variable X is *memoryless* if for all nonnegative x and y,

$$P(X > x + y \mid X > y) = P(X > x) \tag{1.108}$$

A continuous random variable is memoryless if and only if it is exponentially distributed.

Example **1.13.** The time-to-failure distribution of any system is the probability distribution of the random variable T measuring the time the system runs prior to failure. Given T is a continuous random variable with density $f(t)$ and probability distribution function $F(t)$, a system's reliability is characterized by its *reliability function*, which is defined by

$$R(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)\,du$$

$R(t)$ is monotonically decreasing, $R(0) = 1$, and $\lim_{t\to\infty}R(t) = 0$. System reliability is often judged by *mean time to failure* (*MTTF*),

$$E[T] = \int_0^\infty tf(t)\,dt = \int_0^\infty R(t)\,dt$$

The *hazard function* h of a system gives the instantaneous failure rate of the system,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t, T > t)}{P(T > t)\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t)}{P(T > t)\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{R(t)\Delta t}$$

At all points at which $f(t)$ is continuous, $F'(t) = f(t)$. Therefore,

$$h(t) = \frac{f(t)}{R(t)} = -\frac{R'(t)}{R(t)}$$

Consequently,

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

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

From a modeling perspective, a constant hazard function, $h(t) = q$, corresponds to a situation in which wear-in failures have been eliminated and the system has not reached the wear-out stage. Such an assumption is often appropriate for the kind of electronic components used in digital signal processing. From the preceding representations of $R(t)$ and $f(t)$, for $h(t) = q$ the time-to-failure distribution is exponential with $f(t) = qe^{-qt}$, $R(t) = e^{-qt}$, and MTTF $= 1/q$. Since the exponential distribution is memoryless, the probability that the system will function longer than time $t + v$ given that it has functioned for time v is the same as the probability that it will function for time t from the outset: the system's reliability is not modified even though it has been in service for some length of time v. The MTTF is independent of the time the system has already been in operation. ■

1.4.5. Beta Distribution

For $\alpha > 0$, $\beta > 0$, the *beta function* is defined by

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt \tag{1.109}$$

It can be shown that

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{1.110}$$

A random variable X is said to possess a *beta distribution* if it has density

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{1.111}$$

for $0 < x < 1$ and $f(x) = 0$ elsewhere. The beta density takes on various shapes and is therefore useful for modeling many kinds of data distributions. If $\alpha < 1$ and $\beta < 1$, the beta density is U-shaped; if $\alpha < 1$ and $\beta \geq 1$, it is reverse J-shaped; if $\alpha \geq 1$ and $\beta < 1$, it is J-shaped; and if $\alpha > 1$ and $\beta > 1$, it possesses a single maximum. If $\alpha = \beta$, then the graph of the density is symmetric. Some beta densities are shown in Fig. 1.2.

The kth moment of the beta distribution is obtained by applying the moment definition directly to obtain

$$\mu_k' = \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1}\, dx$$

$$= \frac{B(\alpha+k,\beta)}{B(\alpha,\beta)}$$

$$= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)} \tag{1.112}$$

Therefore, $\mu = \alpha(\alpha + \beta)^{-1}$ and $\sigma^2 = \alpha\beta(\alpha + \beta)^{-2}(\alpha + \beta + 1)^{-1}$.

The beta distribution is generalized so as to cover the interval (a, b). The *generalized beta distribution* has density

$$f(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}B(\alpha,\beta)} \tag{1.113}$$

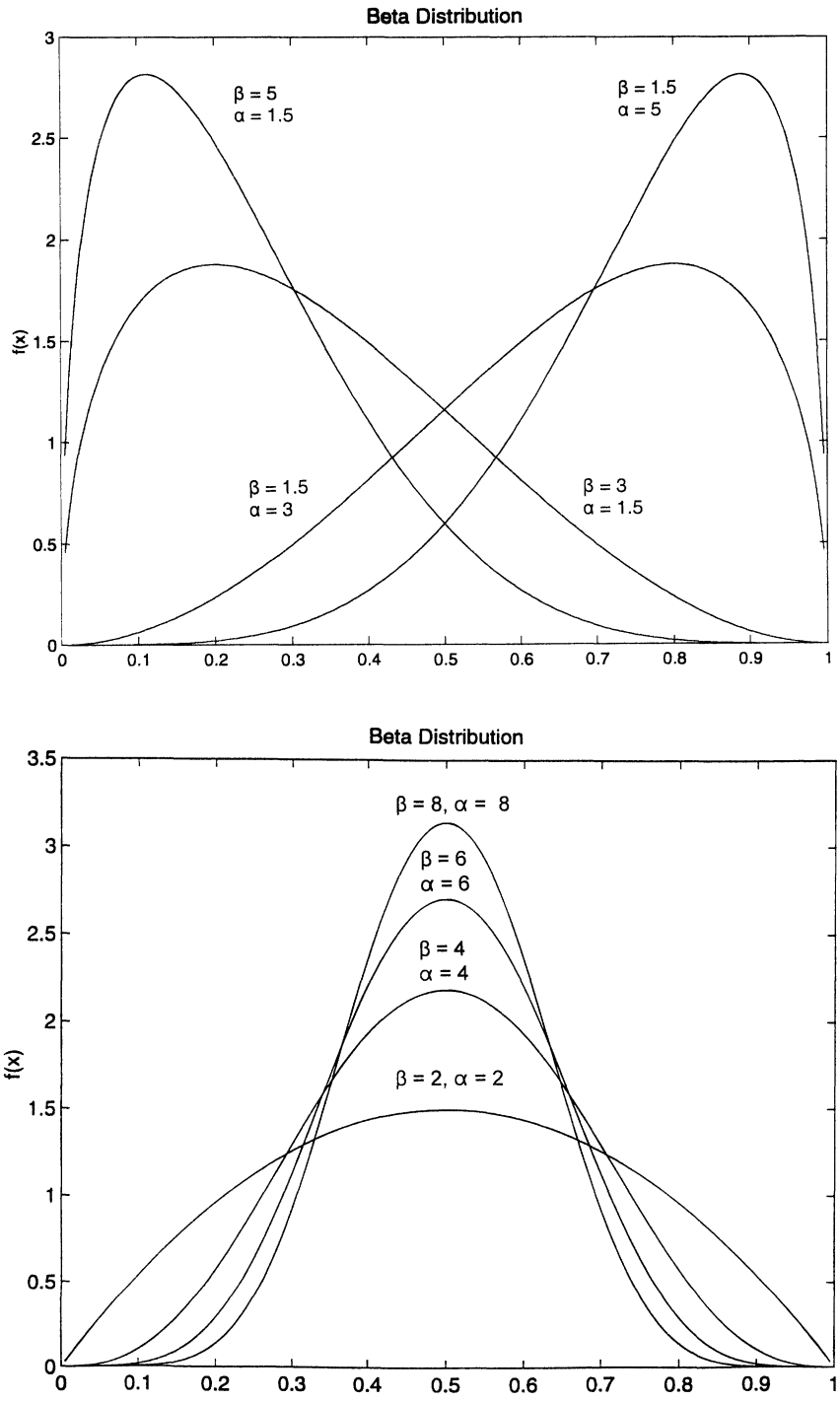

Figure 1.2 Beta densities.

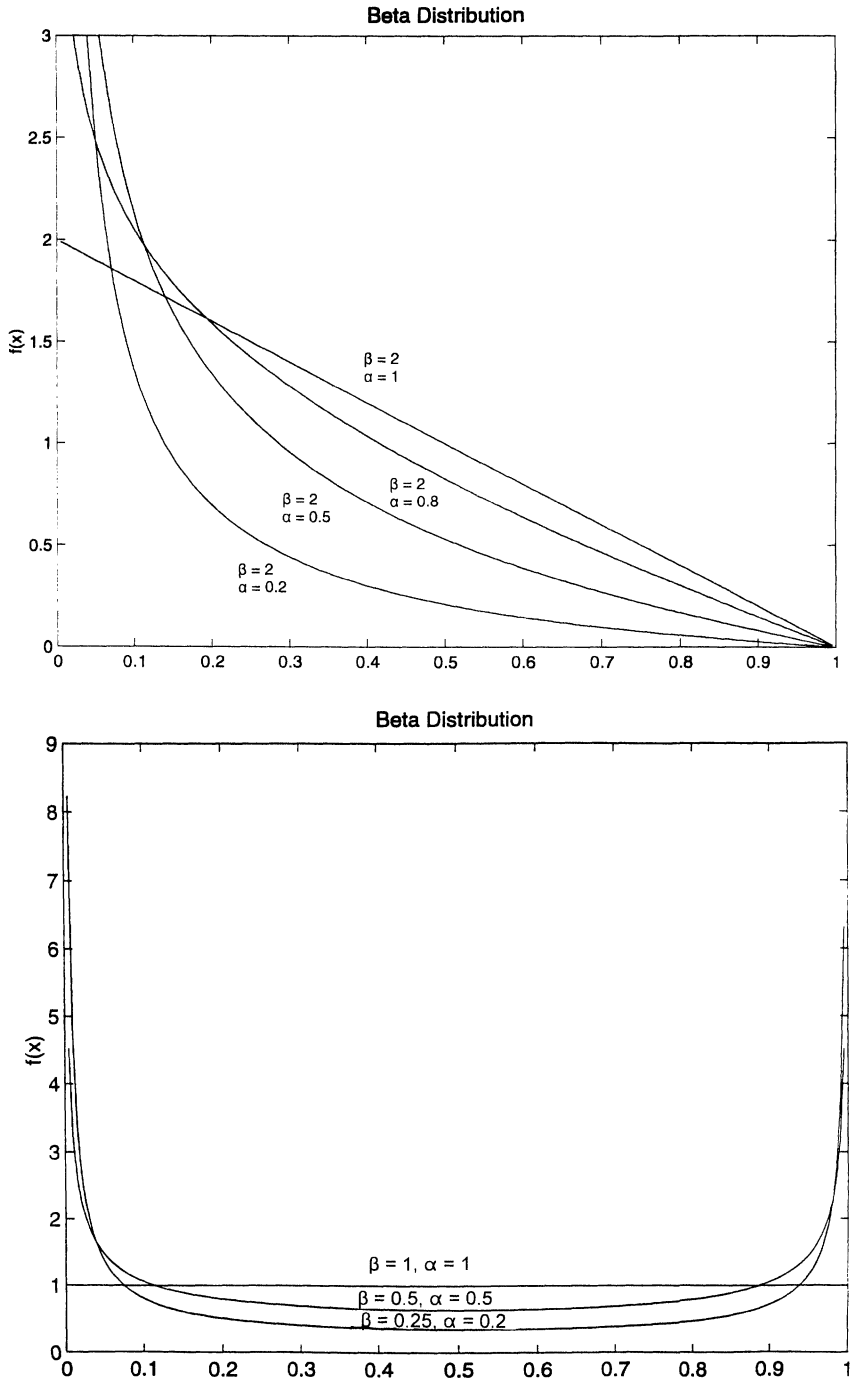

Figure 1.2 (cont): **Beta densities.**

for $a < x < b$, and $f(x) = 0$ elsewhere. Its mean and variance are

$$\mu = a + \frac{(b-a)\alpha}{\alpha+\beta} \tag{1.114}$$

$$\sigma^2 = \frac{(b-a)^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{1.115}$$

Setting $a = 0$ and $b = 1$ gives the mean and variance of the original beta distribution. The uniform distribution introduced in Example 1.7 is a generalized beta distribution with $\alpha = \beta = 1$. It has mean $\mu = (a + b)/2$ and variance $\sigma^2 = (b - a)^2/12$.

1.4.6. Computer Simulation

Processing images and signals involves transformations of random variables. These transformations can be very complex, involving compositions of several highly involved mappings of numerous random variables. Even when there is only one function of a single random variable, it can be difficult to analytically describe the distribution of the output variable in terms of the input distribution. More generally, analytic description is impossible in the majority of cases encountered. Sometimes it is possible to describe some output moments in terms of input moments; however, such descriptions are rarely known for nonlinear operators. Moreover, even if some output moments are known, they may not provide sufficient description of the output distribution. Owing to the intractability of analytic description, it is common to simulate input distributions, operate on the resulting *synthetic* data, and statistically analyze the output data arising from the synthetic input.

The key to the entire procedure is the ability to simulate data whose empirical distribution fits well the theoretical distribution of the random variable of interest. Suppose X is a random variable and $x_1, x_2, \ldots, x_m$ are computer-generated data meant to simulate the behavior of X. If the empirical distribution composed of $x_1, x_2, \ldots, x_m$ is a good approximation of the theoretical distribution, then interval probabilities of the form $P(a < X < b)$ should be well approximated by the proportion of synthetic data in the interval (a, b).

The basis of computer generation of *random values* (outcomes of a random variable) is simulation of the uniformly distributed random variable U over (0, 1). From the perspective that an outcome of U is a nonterminating decimal, generation of U values involves uniformly random selection of digits between 0 and 9, inclusively, and concatenation to form a string of such digits. The result will be a finite decimal expansion and we will take the view that such finite expansions constitute the outcomes of U. Since the strings

are finite, say of length r, the procedure does not actually generate all possible outcomes of U; nonetheless, we must be satisfied with a given degree of approximation. A typical generated value of U takes the form $u = 0.d_1d_2\cdots d_r$, where, for $i = 1, 2,..., r$, d_i is an outcome of a random variable possessing equally likely outcomes between 0 and 9.

The digits $d_1, d_2,..., d_r$ can be generated by uniformly randomly selecting, with replacement, balls numbered 0 through 9 from an urn. In practice, however, the digits are generated by a nonrandom process whose outcomes, called *pseudorandom values*, simulate actual randomness. There are various schemes, called *random-number generators*, for producing pseudorandom values. Not only do they vary in their ability to simulate randomness, but they also require experience to be used effectively.

Besides having routines to generate random values for the uniform distribution, many computer systems also generate random values for the standard normal distribution. Other commonly employed distributions can be simulated by using the random values generated for the uniform distribution.

If F is a continuous probability distribution function that is strictly increasing, then

$$X = F^{-1}(U) \tag{1.116}$$

is a random variable possessing the probability distribution function F. Indeed,

$$F_X(x) = P(F^{-1}(U) \le x) = P(U \le F(x)) = F(x) \tag{1.117}$$

the last equality holding because U is uniformly distributed over (0, 1) and $0 \le F(x) \le 1$.

***Example* 1.14.** To simulate an exponentially distributed random variable X with parameter b, consider its probability distribution function

$$u = F(x) = 1 - e^{-bx}$$

Solving for x in terms of u gives $x = -b^{-1}\log(1 - u)$. According to Eq. 1.116,

$$X = -b^{-1}\log(1 - U)$$

has an exponential distribution with mean $1/b$. To generate exponentially distributed random values, generate uniformly distributed random values and apply the expression of X in terms of U. Since $1 - U$ is uniformly distributed, the representation of X can be simplified to

$X = -b^{-1}\log U$ ■

1.5. Multivariate Distributions

To the extent that phenomena are related, so too are measurements of those phenomena. Since measurements are mathematically treated as random variables, we need to study properties of random variables taken as collections; indeed, in the most general sense the theory of random processes concerns collections of random variables. If $X_1, X_2,..., X_n$ are n random variables, a *random vector* is defined by

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \tag{1.118}$$

Each random variable X_k is a mapping from a sample space into $\Re$; $\mathbf{X}$ is a mapping from the sample space into n-dimensional Euclidean space $\Re^n$. Whereas the probability distributions of $X_1, X_2,..., X_n$ can be determined from the distribution of $\mathbf{X}$, except in special circumstances the distribution of $\mathbf{X}$ cannot be determined from the individual distributions of $X_1, X_2,..., X_n$. To conserve space and to keep matrix-vector algebraic operations consistent, we will write $\mathbf{X} = (X_1, X_2,..., X_n)'$, the prime denoting transpose, so that $\mathbf{X}$ will always be a column vector.

For single random variables we have considered the manner in which a random variable induces a probability measure on the Borel σ-algebra in $\Re$ and how this probability is interpreted as the inclusion probability $P(X \in B)$ for a Borel set $B \subset \Re$. Although we will not go into detail, we note that the *Borel σ-algebra* in $\Re^n$ is the smallest σ-algebra containing all open sets in $\Re^n$, that sets in the Borel σ-algebra are again called *Borel sets*, that all unions and intersections of open and closed sets in $\Re^n$ are Borel sets, and that $\mathbf{X}$ induces a probability measure on the Borel σ-algebra by defining the inclusion probabilities $P(\mathbf{X} \in B)$ for any Borel set $B \subset \Re^n$. We leave the details of product measures and their related σ-algebras to a text on measure theory.

1.5.1. Jointly Distributed Random Variables

If $X_1, X_2,..., X_n$ are n discrete random variables, then their *joint (multivariate) distribution* is defined by the *joint probability mass function*

$$f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \tag{1.119}$$

The joint mass function is nonnegative, there exists a countable set of points $(x_1, x_2,..., x_n) \in \Re^n$ such that $f(x_1, x_2,..., x_n) > 0$, and

$$\sum_{\{(x_1,x_2,...,x_n):f(x_1,x_2,...,x_n)>0\}} f(x_1, x_2, \ldots, x_n) = 1 \tag{1.120}$$

Moreover, for any Borel set $B \subset \Re^n$,

$$P((X_1, X_2,..., X_n)' \in B) = \sum_{\{(x_1,x_2,...,x_n)\in B:f(x_1,x_2,...,x_n)>0\}} f(x_1, x_2, \ldots, x_n) \tag{1.121}$$

The continuous random variables $X_1, X_2,..., X_n$ are said to possess a *multivariate distribution* defined by the *joint density* $f(x_1, x_2,..., x_n) \geq 0$ if, for any Borel set $B \subset \Re^n$,

$$P((X_1, X_2,..., X_n)' \in B) = \iint_B \cdots \int f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n \tag{1.122}$$

If $f(x_1, x_2,..., x_n)$ is a function of n variables such that $f(x_1, x_2,..., x_n) \geq 0$ and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n = 1 \tag{1.123}$$

then there exist continuous random variables $X_1, X_2,..., X_n$ possessing multivariate density $f(x_1, x_2,..., x_n)$. As in the univariate setting, employing delta functions allows us to use continuous-variable notation to represent both continuous and discrete random variables.

***Example* 1.15.** This example treats a discrete multivariate distribution that generalizes the binomial distribution. Consider an experiment satisfying the following three conditions: (1) the experiment consists of n independent trials; (2) for each trial there are r possible outcomes, $w_1, w_2,..., w_r$; (3) there exist numbers $p_1, p_2,..., p_r$ such that, for $j = 1, 2,..., r$, on each trial the probability of outcome w_j is p_j. For $j = 1, 2,..., r$, the random variable X_j counts the number of times that outcome w_j occurs during the n trials. The sample space for the experiment is the set of n-vectors whose components are chosen from the set $\{w_1, w_2,..., w_r\}$ of trial outcomes. Assuming equiprobability, a specific n-vector having x_j components with w_j for $j = 1, 2,..., r$ has probability $p_1^{x_1} p_2^{x_2} \cdots p_r^{x_r}$ of occurring. To obtain the probability of obtaining $X_1 = x_1, X_2 = x_2,..., X_r = x_r$, we need to count the number of such vectors (as in the case of the binomial distribution). Although we will not prove it, the number of n-vectors having x_1 components with w_1, x_2

components with $w_2,\ldots,$ x_r components with w_r is given by the *multinomial coefficient*

$$\binom{n}{x_1, x_2, \ldots, x_r} = \frac{n!}{x_1!\, x_2!\cdots x_r!}$$

Hence, the joint density of $X_1, X_2,\ldots, X_r$ is

$$f(x_1, x_2,\ldots, x_r) = \frac{n!}{x_1!\, x_2!\cdots x_r!}\, p_1^{x_1} p_2^{x_2}\cdots p_r^{x_r}$$

The distribution is known as the *multinomial distribution*. ■

Consider two jointly distributed random variables X and Y. Each has its own univariate distribution and in the context of the joint density $f(x, y)$ the corresponding densities $f_X(x)$ and $f_Y(y)$ are called *marginal densities*. The marginal density for X is derived from the joint density by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \tag{1.124}$$

To see this, note that, for any Borel set $B \subset \mathfrak{R}$,

$$\begin{aligned}\int_B f_X(x)\, dx &= P(X \in B)\\ &= P((X, Y)' \in B \times \mathfrak{R})\\ &= \int_B \int_{-\infty}^{\infty} f(x, y)\, dy dx\end{aligned} \tag{1.125}$$

The marginal density for Y is similarly derived by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx \tag{1.126}$$

More generally, if $X_1, X_2,\ldots, X_n$ possess joint density $f(x_1, x_2,\ldots, x_n)$, then the marginal density for X_k, $k = 1, 2,\ldots, n$, is obtained by

$$f_{X_k}(x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_n \cdots dx_{k+1} dx_{k-1} \cdots dx_1 \tag{1.127}$$

the integral being $(n - 1)$-fold over all variables, excluding x_k. Joint marginal densities can be obtained by integrating over subsets of the variables. For instance, if X, Y, U, and V are jointly distributed, then the joint marginal density for X and Y is given by

$$f_{X,Y}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y,U,V}(x, y, u, v)\, dv du \tag{1.128}$$

Should the variables be discrete, integrals become sums over the appropriate variables.

The *joint probability distribution function* for random variables X and Y is defined by

$$F(x, y) = P(X \le x, Y \le y) \tag{1.129}$$

If X and Y possess the joint density $f(x, y)$, then

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(r, s)\, dr ds \tag{1.130}$$

If the random variables are continuous, then partial differentiation with respect to x and y yields

$$\frac{\partial^2}{\partial x \partial y} F(x, y) = f(x, y) \tag{1.131}$$

at points of continuity of $f(x, y)$. As with univariate distributions, probability distribution functions characterize random variables up to identical distribution.

Theorem 1.10. If X and Y possess the joint probability distribution function $F(x, y)$, then

(i) $\lim_{\min\{x,y\} \to \infty} F(x, y) = 1$

(ii) $\lim_{x \to -\infty} F(x, y) = 0, \lim_{y \to -\infty} F(x, y) = 0$

(iii) $F(x, y)$ is continuous from the right in each variable.

(iv) If $a < b$ and $c < d$, then

$$F(b,d) - F(a,d) - F(b,c) + F(a,c) \geq 0 \tag{1.132}$$

Moreover, if $F(x, y)$ is any function satisfying the four conditions, then there exist random variables X and Y having joint probability distribution function $F(x, y)$. ■

In the bivariate continuous case,

$$P(a < X < b, c < Y < d) = F(b,d) - F(a,d) - F(b,c) + F(a,c) \tag{1.133}$$

Thus, we see the necessity of the condition in Eq. 1.132. Theorem 1.10 can be extended to any finite number of random variables.

1.5.2. Conditioning

Conditioning is a key probabilistic concept for signal processing because it lies at the foundation of filtering systems. If X and Y are discrete and possess joint density $f(x, y)$, then a natural way to define the probability that $Y = y$ given that $X = x$ is by

$$P(Y = y | X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{f(x,y)}{f_X(x)} \tag{1.134}$$

The middle expression is undefined for continuous distributions because the denominator must be zero; however, the last is defined when $f_X(x) > 0$. It is taken as the definition.

If X and Y possess joint density $f(x, y)$, then for all x such that $f_X(x) > 0$, the *conditional density* of Y given $X = x$ is defined by

$$f(y|x) = \frac{f(x,y)}{f_X(x)} \tag{1.135}$$

For a given x, $f(y \mid x)$ is a function of y and is a legitimate density because

$$\int_{-\infty}^{\infty} f(y|x)\,dy = \frac{1}{f_X(x)} \int_{-\infty}^{\infty} f(x,y)\,dy = 1 \tag{1.136}$$

The random variable associated with the conditional density is called the *conditional random variable* Y given x and is denoted by $Y|x$. It has an expectation called the *conditional expectation* (*conditional mean*), denoted by $E[Y \mid x]$ or $\mu_{Y|x}$, and defined by

$$E[Y \mid x] = \int_{-\infty}^{\infty} y f(y|x)\, dy \tag{1.137}$$

It also has a conditional variance,

$$\mathrm{Var}[Y \mid x] = E[(Y|x - \mu_{Y|x})^2] \tag{1.138}$$

Conditioning can be extended to $n + 1$ random variables $X_1, X_2, \ldots, X_n, Y$. If the joint densities for $X_1, X_2, \ldots, X_n, Y$ and $X_1, X_2, \ldots, X_n$ are $f(x_1, x_2, \ldots, x_n, y)$ and $f(x_1, x_2, \ldots, x_n)$, respectively, then the conditional density of Y given $X_1, X_2, \ldots, X_n$ is defined by

$$f(y|x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n, y)}{f(x_1, x_2, \ldots, x_n)} \tag{1.139}$$

for all $(x_1, x_2, \ldots, x_n)$ such that $f(x_1, x_2, \ldots, x_n) > 0$. The conditional expectation and variance of Y given $x_1, x_2, \ldots, x_n$ are defined via the conditional density. In particular,

$$E[Y|x_1, x_2, \ldots, x_n] = \int_{-\infty}^{\infty} y f(y|x_1, x_2, \ldots, x_n)\, dy \tag{1.140}$$

The conditional expectation plays a major role in filter optimization.

***Example* 1.16.** Random variables X and Y are said to possess a *joint uniform distribution* over region $R \subset \Re^2$ if their joint density $f(x, y)$ is defined by $f(x, y) = 1/v[R]$ for $(x, y) \in R$ and $f(x, y) = 0$ for $(x, y) \notin R$, where $v[R]$ is the area of R (assuming $v[R] > 0$). Let X and Y be jointly uniformly distributed over the triangular region R consisting of the portion of the plane bounded by the x axis, the line $x = 1$, and the line $y = x$. For $0 \le x \le 1$,

$$f_X(x) = \int_0^x 2\, dy = 2x$$

and $f_X(x) = 0$ for $x \notin [0, 1]$. For $0 \le y \le 1$,

$$f_Y(y) = \int_y^1 2\,dx = 2(1-y)$$

and $f_Y(y) = 0$ for $y \notin [0, 1]$. The conditional density for $Y|x$ is defined for $0 < x \leq 1$ and is given by

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{2}{2x} = \frac{1}{x}$$

for $0 \leq y \leq x$ and $f(y \mid x) = 0$ for $y \notin [0, x]$. The conditional density for $X|y$ is defined for $0 \leq y < 1$ and is given by

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{2}{2(1-y)} = \frac{1}{1-y}$$

for $y \leq x \leq 1$ and $f(x \mid y) = 0$ for $x \notin [y, 1]$. Thus, $Y|x$ and $X|y$ are uniformly distributed over $[0, x]$ and $[y, 1]$, respectively. Consequently, $E[Y \mid x] = x/2$ and $E[X \mid y] = (1 + y)/2$. These conditional expectations can be obtained by integration (Eq. 1.137). ■

1.5.3. Independence

Cross multiplication in Eq. 1.135 yields

$$f(x,y) = f_X(x) f(y \mid x) \tag{1.141}$$

If

$$f(x,y) = f_X(x) f_Y(y) \tag{1.142}$$

then Eq. 1.141 reduces to $f(y \mid x) = f_Y(y)$, so that conditioning by x does not affect the probability distribution of Y. If Eq. 1.142 holds, then X and Y are said to be *independent*. In general, random variables $X_1, X_2, \ldots, X_n$ possessing multivariate density $f(x_1, x_2, \ldots, x_n)$ are said to be *independent* if

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) \tag{1.143}$$

Otherwise they are *dependent*.

If $X_1, X_2, \ldots, X_n$ are independent, then so too is any subset of $X_1, X_2, \ldots, X_n$. To see this for three jointly distributed continuous random variables X, Y, Z, note that

$$
\begin{aligned}
f_{X,Y}(x,y) &= \int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,dz \\
&= \int_{-\infty}^{\infty} f_X(x) f_Y(y) f_Z(z)\,dz \\
&= f_X(x) f_Y(y) \int_{-\infty}^{\infty} f_Z(z)\,dz \\
&= f_X(x) f_Y(y)
\end{aligned}
\tag{1.144}
$$

If $X_1, X_2, \ldots, X_n$ are independent, then for any Borel sets $B_1, B_2, \ldots, B_n$,

$$
\begin{aligned}
P\left(\bigcap_{i=1}^{n} (X_i \in B_i)\right) &= P((X_1, X_2, \ldots, X_n) \in B_1 \times B_2 \times \cdots \times B_n) \\
&= \int_{B_1}\int_{B_2}\cdots\int_{B_n} f(x_1, x_2, \ldots, x_n)\,dx_n dx_{n-1}\cdots dx_1 \\
&= \int_{B_1}\int_{B_2}\cdots\int_{B_n} f_{X_1}(x_1) f_{X_2}(x_2)\cdots f_{X_n}(x_n)\,dx_n dx_{n-1}\cdots dx_1 \\
&= \prod_{i=1}^{n} \int_{B_i} f_{X_i}(x_i)\,dx_i \\
&= \prod_{i=1}^{n} P(X_i \in B_i)
\end{aligned}
\tag{1.145}
$$

In general, given the multivariate density for a finite collection of random variables, it is possible to derive the marginal densities via integration according to Eq. 1.127. On the contrary, it is not generally possible to obtain the multivariate density from the marginal densities; however, if the random variables are independent, then the multivariate density can be so derived. Many experimental designs postulate independence just so the joint density can be expressed as a product of the marginal densities.

Example 1.17. If $X_1, X_2,..., X_n$ are independent normally distributed random variables with means $\mu_1, \mu_2,..., \mu_n$ and standard deviations $\sigma_1, \sigma_2,..., \sigma_n$, respectively, then the multivariate density for $X_1, X_2,..., X_n$ is given by

$$f(x_1, x_2,..., x_n) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x_k-\mu_k}{\sigma_k}\right)^2}$$

$$= \frac{1}{(2\pi)^{n/2}} \left(\prod_{k=1}^{n} \sigma_k^{-1}\right) \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{x_k-\mu_k}{\sigma_k}\right)^2\right)\right]$$

$$= \frac{1}{\sqrt{(2\pi)^n \det[\mathbf{K}]}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

where $\mathbf{x} = (x_1, x_2,..., x_n)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2,..., \mu_n)'$, and $\mathbf{K}$ is the diagonal matrix whose diagonal contains the variances of $X_1, X_2,..., X_n$,

$$\mathbf{K} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

■

1.6. Functions of Several Random Variables

When dealing with random processes we are rarely concerned with a function of a single random variable; for most systems there are several random inputs and one or more random outputs. Given n input random variables $X_1, X_2,..., X_n$ to a system, if the system outputs a single numerical value, then the output is a function of the form

$$Y = g(X_1, X_2,..., X_n) \tag{1.146}$$

We would like to describe the probability distribution of the output given the joint distribution of the inputs, or at least describe some of the output moments. Typically, this problem is very difficult.

1.6.1. Basic Arithmetic Functions of Two Random Variables

For some basic arithmetic functions of two continuous random variables X and Y having joint density $f(x, y)$, output densities can be found by

differentiation of the probability distribution functions. The probability distribution function of the sum $X + Y$ is given by

$$F_{X+Y}(z) = P(X+Y \le z)$$

$$= \iint_{\{(x,y):x+y\le z\}} f(x,y)\,dxdy$$

$$= \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} f(x,y)\,dy$$

$$= \int_{-\infty}^{\infty} dx \int_{-\infty}^{z} f(x,u-x)\,du \tag{1.147}$$

the last integral resulting from the substitution $u = x + y$. Differentiation yields the density

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f(x,z-x)\,dx = \int_{-\infty}^{\infty} f(z-y,y)\,dy \tag{1.148}$$

where the second integral follows by symmetry. If X and Y are independent, then

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx = \int_{-\infty}^{\infty} f_X(z-x) f_Y(x)\,dx \tag{1.149}$$

which is the convolution of the densities.

The probability distribution function of the product XY is given by

$$F_{XY}(z) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{z/x} f(x,y)\,dy$$

$$= \int_{-\infty}^{\infty} \frac{dx}{|x|} \int_{-\infty}^{z} f\left(x, \frac{u}{x}\right) du \tag{1.150}$$

the last integral resulting from the substitution $u = xy$. Differentiation yields the density

$$f_{XY}(z) = \int_{-\infty}^{\infty} f\left(x, \frac{z}{x}\right) \frac{dx}{|x|} = \int_{-\infty}^{\infty} f\left(\frac{z}{y}, y\right) \frac{dy}{|y|} \tag{1.151}$$

The probability distribution function of the quotient Y/X is given by

$$F_{Y/X}(z) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{xz} f(x, y)\, dy$$

$$= \int_{-\infty}^{\infty} |x| dx \int_{-\infty}^{z} f(x, ux)\, du \tag{1.152}$$

the last integral resulting from the substitution $u = y/x$. Differentiation yields the density

$$f_{Y/X}(z) = \int_{-\infty}^{\infty} f(x, zx)|x| dx = \int_{-\infty}^{\infty} f(zy, y)|y| dy \tag{1.153}$$

Example 1.18. For a quotient $Z = Y/X$ of two independent standard normal random variables, Eq. 1.153 yields

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(zx)|x| dx$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-(zx)^2/2} |x| dx$$

$$= \frac{1}{\pi} \int_{0}^{\infty} e^{-(1+z^2)x^2/2} x dx$$

$$= \frac{1}{\pi(1+z^2)}$$

which is known as the *Cauchy density*. This density does not possess an expectation since

$$\int_{-\infty}^{\infty} |z| f_Z(z)\, dz = \int_{-\infty}^{\infty} \frac{|z|}{\pi(1+z^2)}\, dz$$

does not converge when evaluated by improper integration. ■

The probability distribution function for $Z = (X^2 + Y^2)^{1/2}$, the Euclidean norm of the random vector $(X, Y)'$, is given by $F_Z(z) = 0$ for $z < 0$ and, for $z \geq 0$,

$$F_Z(z) = \iint_{\{(x,y):x^2+y^2 \leq z^2\}} f(x,y)\, dxdy$$

$$= \int_0^{2\pi} d\theta \int_0^z f(r\cos\theta, r\sin\theta) r\, dr \tag{1.154}$$

where we have transformed to polar coordinates. Differentiation yields

$$f_Z(z) = z \int_0^{2\pi} f(z\cos\theta, z\sin\theta)\, d\theta \tag{1.155}$$

for $z \geq 0$, and $f_Z(z) = 0$ for $z < 0$.

***Example* 1.19.** If X and Y are independent zero-mean, normal random variables possessing a common variance σ^2, then, according to Eq. 1.155, the density of the Euclidean norm Z of $(X, Y)'$ is given by

$$f_Z(z) = z \int_0^{2\pi} f_X(z\cos\theta) f_Y(z\sin\theta)\, d\theta$$

$$= \frac{z}{2\pi\sigma^2} \int_0^{2\pi} e^{-\frac{z^2}{2\sigma^2}}\, d\theta$$

$$= \frac{z}{\sigma^2} e^{-\frac{z^2}{2\sigma^2}}$$

for $z \geq 0$ and $f_Z(z) = 0$ for $z < 0$, which is known as a *Rayleigh density*. ■

1.6.2. Distributions of Sums of Independent Random Variables

An important special case of a function of several random variables $X_1, X_2,\ldots, X_n$ occurs when the function of the random variables is linear,

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \tag{1.156}$$

where $a_1, a_2,\ldots, a_n$ are constants. If $X_1, X_2,\ldots, X_n$ are independent, then we can employ moment-generating functions to discover output distributions for linear functions.

If $X_1, X_2,\ldots, X_n$ are independent random variables possessing joint density $f(x_1, x_2,\ldots, x_n)$ and their moment-generating functions exist for $t < t_0$, then the moment-generating function of

$$Y = X_1 + X_2 + \cdots + X_n \tag{1.157}$$

exists for $t < t_0$ and is given by

$$\begin{aligned} M_Y(t) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left[t\sum_{k=1}^{n}x_k\right]f(x_1,x_2,\ldots,x_n)\,dx_1dx_2\cdots dx_n \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{k=1}^{n}e^{tx_k}f(x_k)\,dx_1dx_2\cdots dx_n \\ &= \prod_{k=1}^{n}M_{X_k}(t) \end{aligned} \tag{1.158}$$

where the second equality follows from the properties of exponentials and the independence of the random variables. The relation of Eq. 1.158 has been demonstrated for continuous random variables; the demonstration is similar for discrete random variables, except that integrals are replaced by sums.

Example 1.20. Suppose $X_1, X_2,\ldots, X_n$ are independent gamma-distributed random variables with X_k having parameters α_k and β, for $k = 1, 2,\ldots, n$. Since the moment-generating function of X_k is $(1-\beta t)^{-\alpha_k}$, Eq. 1.158 shows the moment-generating function of the sum of $X_1, X_2,\ldots, X_n$ to be

$$M_Y(t) = \prod_{k=1}^{n}(1-\beta t)^{-\alpha_k} = (1-\beta t)^{-(\alpha_1+\alpha_2+\cdots+\alpha_n)}$$

which is the moment-generating function of a gamma-distributed random variable with parameters $\alpha_1 + \alpha_2 + \cdots + \alpha_n$ and β. By uniqueness, the sum Y is such a variable. As a special case, the sum of n identically distributed exponential random variables with parameter b is gamma distributed with parameters $\alpha = n$ and $\beta = 1/b$. Such a random variable is also said to be n-Erlang with parameter $1/b$.

If $U_1, U_2, \ldots, U_n$ are independent uniform random variables on (0, 1), then, from Example 1.14, we know that $-b^{-1}\log U_k$ is exponentially distributed with parameter b, for $k = 1, 2, \ldots, n$. Hence

$$X = -b^{-1}\sum_{k=1}^{n} \log U_k$$

is gamma distributed with $\alpha = n$ and $\beta = 1/b$. X can be simulated via computer generation of independent uniform random variables. ■

***Example* 1.21.** Suppose $X_1, X_2, \ldots, X_n$ are independent Poisson-distributed random variables with X_k having mean λ_k, for $k = 1, 2, \ldots, n$. Since the moment-generating function of X_k is $\exp[\lambda_k(e^t - 1)]$, Eq. 1.158 shows the moment-generating function of the sum of $X_1, X_2, \ldots, X_n$ to be

$$M_Y(t) = \prod_{k=1}^{n} \exp[\lambda_k (e^t - 1)] = \exp\left[(e^t - 1)\sum_{k=1}^{n} \lambda_k\right]$$

which is the moment-generating function for a Poisson-distributed random variable with mean $\lambda_1 + \lambda_2 + \cdots + \lambda_n$. By uniqueness, the sum Y is such a variable. ■

***Example* 1.22.** Suppose $X_1, X_2, \ldots, X_n$ are independent normally distributed random variables such that, for $k = 1, 2, \ldots, n$, X_k has mean μ_k and variance σ_k^2. Since the moment-generating function of X_k is $\exp[\mu_k t + \sigma_k^2 t^2/2]$, Eq. 1.158 shows the moment-generating function of the linear combination of Eq. 1.156 to be

$$M_Y(t) = \prod_{k=1}^{n} M_{X_k}(a_k t)$$

$$= \prod_{k=1}^{n} \exp\left[a_k \mu_k t + \frac{a_k^2 \sigma_k^2 t^2}{2}\right]$$

$$= \exp\left[\left(\sum_{k=1}^{n} a_k \mu_k\right) t + \left(\sum_{k=1}^{n} a_k^2 \sigma_k^2\right) \frac{t^2}{2}\right]$$

Hence, Y is normally distributed with mean and variance

$$\mu_Y = \sum_{k=1}^{n} a_k \mu_k$$

$$\sigma_Y^2 = \sum_{k=1}^{n} a_k^2 \sigma_k^2$$ ■

1.6.3. Joint Distributions of Output Random Variables

Many systems have several output random variables as well as several input variables and for these it is desirable (if possible) to have the joint distribution of the output random variables in terms of the distribution of the input variables.

For the case of two discrete input random variables, X and Y, and two discrete output random variables, U and V, there exist functions g and h such that

$$U = g(X, Y) \tag{1.159}$$

$$V = h(X, Y)$$

and the output probability mass function is

$$f_{U,V}(u, v) = P(g(X,Y) = u, h(X,Y) = v)$$

$$= \sum_{\{(x,y):g(x,y)=u,h(x,y)=v\}} f_{X,Y}(x, y) \tag{1.160}$$

Now suppose the vector mapping

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} g(x,y) \\ h(x,y) \end{pmatrix} \tag{1.161}$$

is one-to-one and has the inverse vector mapping

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r(u,v) \\ s(u,v) \end{pmatrix} \tag{1.162}$$

Then

$$\{(x, y): g(x, y) = u, h(x, y) = v\} = \{(r(u, v), s(u, v))\} \tag{1.163}$$

and Eq. 1.160 reduces to

$$f_{U,V}(u, v) = f_{X,Y}(r(u, v), s(u, v)) \tag{1.164}$$

The analysis extends to any finite-dimensional vector mapping.

***Example* 1.23.** Suppose X and Y are independent binomially distributed random variables, X having parameters n and p, and Y having parameters m and d. Their joint density is the product of their individual densities,

$$f_{X,Y}(x,y) = \binom{n}{x}\binom{m}{y} p^x d^y (1-p)^{n-x} (1-d)^{m-y}$$

Suppose the output random variables are defined by $U = X + Y$ and $V = X - Y$. The vector mapping

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} g(x,y) \\ h(x,y) \end{pmatrix} = \begin{pmatrix} x+y \\ x-y \end{pmatrix}$$

possesses the unique solution

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r(u,v) \\ s(u,v) \end{pmatrix} = \begin{pmatrix} (u+v)/2 \\ (u-v)/2 \end{pmatrix}$$

From Eq. 1.164,

$$f_{U,V}(u,v) = \binom{n}{(u+v)/2}\binom{m}{(u-v)/2} p^{(u+v)/2} d^{(u-v)/2} (1-p)^{n-(u+v)/2} (1-d)^{m-(u-v)/2}$$

Constraints on the variables u and v are determined by the constraints on x and y. ■

Theorem 1.11. Suppose the continuous random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ possesses multivariate density $f(x_1, x_2, \ldots, x_n)$ and the vector mapping $\mathbf{x} \to \mathbf{u}$ is defined by

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} g_1(x_1, x_2, \ldots, x_n) \\ g_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ g_n(x_1, x_2, \ldots, x_n) \end{pmatrix} \tag{1.165}$$

where $g_1, g_2, \ldots, g_n$ have continuous partial derivatives and the mapping is one-to-one on the set $A_{\mathbf{x}}$ of all vectors $\mathbf{x}$ such that $f(\mathbf{x}) > 0$. If $A_{\mathbf{u}}$ denotes the set of all vectors corresponding to vectors in $A_{\mathbf{x}}$ and if the inverse vector mapping $\mathbf{u} \to \mathbf{x}$ is defined by

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} r_1(u_1, u_2, \ldots, u_n) \\ r_2(u_1, u_2, \ldots, u_n) \\ \vdots \\ r_n(u_1, u_2, \ldots, u_n) \end{pmatrix} \tag{1.166}$$

then the joint density of the random vector

$$\mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix} = \begin{pmatrix} g_1(X_1, X_2, \ldots, X_n) \\ g_2(X_1, X_2, \ldots, X_n) \\ \vdots \\ g_n(X_1, X_2, \ldots, X_n) \end{pmatrix} \tag{1.167}$$

is given by

$$f_{\mathbf{U}}(\mathbf{u}) = \begin{cases} f_{\mathbf{X}}(r_1(\mathbf{u}), r_2(\mathbf{u}), \ldots, r_n(\mathbf{u}))|J(\mathbf{x}; \mathbf{u})|, & \text{if } \mathbf{u} \in A_{\mathbf{u}} \\ 0, & \text{otherwise} \end{cases} \tag{1.168}$$

where $J(\mathbf{x}; \mathbf{u})$, the *Jacobian* of the mapping $\mathbf{u} \to \mathbf{x}$, is defined by the determinant

$$J(\mathbf{x}; \mathbf{u}) = \det \begin{bmatrix} \partial x_1/\partial u_1 & \partial x_1/\partial u_2 & \cdots & \partial x_1/\partial u_n \\ \partial x_2/\partial u_1 & \partial x_2/\partial u_2 & \cdots & \partial x_2/\partial u_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial x_n/\partial u_1 & \partial x_n/\partial u_2 & \cdots & \partial x_n/\partial u_n \end{bmatrix} \tag{1.169}$$

■

***Example* 1.24.** As in Example 1.23, consider the vector mapping $U = X + Y$, $V = X - Y$, but now let X and Y be jointly uniformly distributed over the unit square $A_x = (0, 1)^2$. A_u is determined by solving the variable constraint pair

$$0 < \frac{u+v}{2} < 1, \qquad 0 < \frac{u-v}{2} < 1$$

Since $\partial x/\partial u = \partial x/\partial v = \partial y/\partial u = 1/2$ and $\partial y/\partial v = -1/2$, the Jacobian is

$$J(\mathbf{x}; \mathbf{u}) = \det\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix} = -\frac{1}{2}$$

By Theorem 1.11, $f_{U,V}(u, v) = 1/2$ for $(u, v) \in A_u$ and $f_{U,V}(u, v) = 0$ otherwise. ■

1.6.4. Expectation of a Function of Several Random Variables

For taking the expectation of a function of several random variables, there exists the following extension of Theorem 1.7. An important consequence of the new theorem is that the expected-value operator is a linear operator.

Theorem 1.12. Suppose $X_1, X_2, \ldots, X_n$ have joint density $f(x_1, x_2, \ldots, x_n)$ and $g(x_1, x_2, \ldots, x_n)$ is a piecewise continuous function of $x_1, x_2, \ldots, x_n$. Then

$$E[g(X_1, X_2, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n \tag{1.170}$$

where the integral is n-fold. ■

For discrete random variables there is no restriction on $g(x_1, x_2, \ldots, x_n)$ and the integral representation of Eq. 1.170 becomes

$$E[g(X_1, X_2, \ldots, X_n)] = \sum_{\{(x_1, x_2, \ldots, x_n): f(x_1, x_2, \ldots, x_n) > 0\}} g(x_1, x_2, \ldots, x_n) f(x_1, x_2 \ldots, x_n) \tag{1.171}$$

Theorem 1.13. For any random variables $X_1, X_2, \ldots, X_n$ possessing expectations and constants $a_1, a_2, \ldots, a_n$,

$$E\left[\sum_{k=1}^{n} a_k X_k\right] = \sum_{k=1}^{n} a_k E[X_k] \qquad \blacksquare \tag{1.172}$$

Linearity is demonstrated by

$$E\left[\sum_{k=1}^{n} a_k X_k\right] = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sum_{k=1}^{n} a_k x_k\right) f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n$$

$$= \sum_{k=1}^{n} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} a_k x_k f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n$$

$$= \sum_{k=1}^{n} a_k \int_{-\infty}^{\infty} x_k f_{X_k}(x_k)\, dx_k \tag{1.173}$$

$$= \sum_{k=1}^{n} a_k E[X_k]$$

1.6.5. Covariance

Moments can be generalized to collections of jointly distributed random variables. Key to the analysis of linear systems and random processes is the covariance. We are mainly concerned with bivariate moments.

Given two random variables X and Y and integers $p \geq 0$ and $q \geq 0$, the $(p + q)$-order *product moment* is defined by

$$\mu'_{pq} = E[X^p Y^q] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x, y)\, dydx \tag{1.174}$$

The $(p + q)$-order *central moment* is

$$\mu_{pq} = E[(X - \mu_X)^p (Y - \mu_Y)^q]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)^p (y - \mu_Y)^q f(x, y)\, dydx \tag{1.175}$$

In both cases we assume that the defining integral is absolutely convergent.

The second-order central product moment μ_{11} is called the *covariance* and is given by

$$\text{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] \tag{1.176}$$

We sometimes denote the covariance of X and Y by σ_{XY}^2. If it exists, the covariance is conveniently expressed as

$$\text{Cov}[X, Y] = E[XY] - \mu_X\mu_Y \tag{1.177}$$

which can be seen by expanding the integral expression for the covariance.

If X and Y are independent, then

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dydx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)\,dydx \\ &= \int_{-\infty}^{\infty} xf_X(x)\,dx \int_{-\infty}^{\infty} yf_Y(y)\,dy \\ &= E[X]E[Y] \end{aligned} \tag{1.178}$$

In particular, from Eq. 1.177, if X and Y are independent, then $\text{Cov}[X, Y] = 0$.

The covariance provides a measure of the linear relationship between random variables; however, the deviations $X - \mu_X$ and $Y - \mu_Y$ are dependent on the units in which X and Y are measured. A normalized measure is given by the *correlation coefficient*

$$\rho_{XY} = \frac{\text{Cov}[X,Y]}{\sigma_X\sigma_Y} \tag{1.179}$$

If $\rho_{XY} = 0$, then the random variables are said to be *uncorrelated*. If the variables are independent, then according to Eq. 1.178 they are uncorrelated. The converse, however, is not valid: uncorrelated variables need not be independent. The next theorem specifies the manner in which the correlation coefficient provides a measure of linearity between random variables. It is simply a statement of the Schwarz inequality in terms of the correlation coefficient.

Theorem 1.14. For any random variables X and Y,

$$-1 \le \rho_{XY} \le 1 \tag{1.180}$$

and $|\rho_{XY}| = 1$ if and only if there exist constants $a \neq 0$ and b such that

$$P(Y = aX + b) = 1 \quad \blacksquare \tag{1.181}$$

According to the theorem, the probability mass of the joint distribution lies on a straight line if and only if $|\rho_{XY}| = 1$. The correlation coefficient lies between -1 and $+1$. The closer it is to either extreme, the more the mass is linearly concentrated. For $|\rho_{XY}| = 1$, Y is (up to possibly a set of probability zero) a linear function of X.

Theorem 1.15. If Y is given by the linear combination of the random variables $X_1, X_2, \ldots, X_n$ in Eq. 1.156, then

$$\text{Var}[Y] = \sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \text{Cov}[X_j, X_k] \tag{1.182}$$

If $X_1, X_2, \ldots, X_n$ are uncorrelated, then all covariance terms vanish except for $j = k$ and

$$\text{Var}[Y] = \sum_{k=1}^{n} a_k^2 \text{Var}[X_k] \quad \blacksquare \tag{1.183}$$

Equation 1.182 is demonstrated in the following manner:

$$\begin{aligned}
\text{Var}[Y] &= E[Y^2] - E[Y]^2 \\
&= E\left[\left(\sum_{k=1}^{n} a_k X_k\right)^2\right] - \left(\sum_{k=1}^{n} a_k E[X_k]\right)^2 \\
&= E\left[\sum_{k=1}^{n}\sum_{j=1}^{n} a_k a_j X_k X_j\right] - \sum_{k=1}^{n}\sum_{j=1}^{n} a_k a_j E[X_k]E[X_j] \\
&= \sum_{k=1}^{n}\sum_{j=1}^{n} a_k a_j (E[X_k X_j] - E[X_k]E[X_j]) \qquad (1.184) \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \text{Cov}[X_j, X_k]
\end{aligned}$$

the last equality following from Eq. 1.177.

The *mean vector* and *covariance matrix* for the random vector $\mathbf{X} = (X_1, X_2,..., X_n)'$ are defined by

$$\boldsymbol{\mu} = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix} \tag{1.185}$$

and

$$\mathbf{K} = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{pmatrix} \tag{1.186}$$

where $\sigma_{ij}^2 = \text{Cov}[X_i, X_j]$. The diagonal elements of $\mathbf{K}$ are the variances of X_1, $X_2,..., X_n$.

The covariance matrix $\mathbf{K}$ is real and symmetric. It is *nonnegative definite*, meaning that $\mathbf{v}'\mathbf{K}\mathbf{v} \geq 0$ for any n-vector $\mathbf{v}$. This is shown by the inequality

$$\begin{aligned} 0 &\leq E[|\mathbf{v}'(\mathbf{X} - \boldsymbol{\mu})|^2] \\ &= \mathbf{v}'E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})']\mathbf{v} \\ &= \mathbf{v}'\mathbf{K}\mathbf{v} \end{aligned} \tag{1.187}$$

If $\mathbf{K}$ has eigenvalues $\lambda_1, \lambda_2,..., \lambda_n$, since $\mathbf{K}$ is real and symmetric, it has mutually orthogonal unit-length eigenvectors $\mathbf{e}_1, \mathbf{e}_2,..., \mathbf{e}_n$ corresponding to $\lambda_1, \lambda_2,..., \lambda_n$. If

$$\mathbf{E} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \cdots \ \mathbf{e}_n] \tag{1.188}$$

is the matrix whose columns are $\mathbf{e}_1, \mathbf{e}_2,..., \mathbf{e}_n$, then $\mathbf{E}$ is *unitary*, meaning $\mathbf{E}^{-1} = \mathbf{E}'$, and

$$\mathbf{E}^{-1}\mathbf{K}\mathbf{E} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \tag{1.189}$$

meaning that **K** is *similar* to the diagonal matrix composed of the eigenvalues of **K** (in order down the diagonal). Finally, since **K** is real and symmetric, it is *positive definite*, meaning $\mathbf{v}'\mathbf{K}\mathbf{v} > 0$ for any nonzero n-vector **v**, if and only if all eigenvalues are positive.

Covariance functions of random functions play important roles in the theory and application of random processes. We will have much to say about the covariance matrix in the context of finite random processes; however, we have introduced it here and discussed a few of its linear-algebraic properties for completeness, in particular, because we wish to introduce the multivariate normal distribution.

1.6.6. Multivariate Normal Distribution

To define the multivariate normal distribution, let **K** be a positive definite, real symmetric matrix and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ be an arbitrary vector. A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ is said to have a *multivariate normal (Gaussian) distribution* if it possesses the multivariate density

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det[\mathbf{K}]}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{1.190}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$. When $n = 1$, $\mathbf{K} = (\sigma^2)$, $\det[\mathbf{K}] = \sigma^2$, and $\boldsymbol{\mu}$ is the mean.

The following properties hold for the multivariate normal distribution: (1) $\boldsymbol{\mu}$ is the mean vector; (2) **K** is the covariance matrix; (3) $X_1, X_2, \ldots, X_n$ are independent if and only if **K** is diagonal (which is the case considered in Example 1.17); and (4) the marginal densities are normally distributed.

In the special case where $n = 2$,

$$\mathbf{K} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \tag{1.191}$$

and the density is given by

$$f(x,y) = \frac{\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \tag{1.192}$$

where the marginal random variables X and Y are normally distributed and μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ are the mean of X, mean of Y, variance of X, variance of

Y, and correlation coefficient, respectively. These properties can be demonstrated by performing the appropriate integrations. It follows immediately from the form of the joint density that X and Y are independent if and only if they are uncorrelated (whereas, generally, uncorrelatedness does not imply independence).

In Example 1.22 we have seen that a linear combination of independent normally distributed random variables is normally distributed and we have found the mean and variance of the linear combination. Now consider the more general situation where $\mathbf{X}$ is a random vector possessing an n-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$, $\mathbf{A}$ is a nonsingular $n \times n$ matrix, and $\mathbf{U} = \mathbf{AX}$ is the output random vector. Using Theorem 1.11, we demonstrate that $\mathbf{U}$ possesses a multivariate normal distribution with mean vector $\mathbf{A}\boldsymbol{\mu}$ and covariance matrix $\mathbf{AKA}'$. The transformation $\mathbf{u} = \mathbf{Ax}$ is inverted by $\mathbf{x} = \mathbf{A}^{-1}\mathbf{u}$. Hence, the Jacobian of the mapping $\mathbf{u} \to \mathbf{x}$ is

$$J(\mathbf{x}; \mathbf{u}) = \det[\mathbf{A}^{-1}] = \det[\mathbf{A}]^{-1} \tag{1.193}$$

and, according to Eqs. 1.168 and 1.190,

$$\begin{aligned} f_{\mathbf{U}}(\mathbf{u}) &= \frac{1}{|\det[\mathbf{A}]|\sqrt{(2\pi)^n \det[\mathbf{K}]}} \exp\left[-\frac{1}{2}(\mathbf{A}^{-1}\mathbf{u}-\boldsymbol{\mu})'\mathbf{K}^{-1}(\mathbf{A}^{-1}\mathbf{u}-\boldsymbol{\mu})\right] \\ &= \frac{1}{\sqrt{(2\pi)^n \det[\mathbf{AKA}']}} \exp\left[-\frac{1}{2}(\mathbf{u}-\mathbf{A}\boldsymbol{\mu})'(\mathbf{AKA}')^{-1}(\mathbf{u}-\mathbf{A}\boldsymbol{\mu})\right] \end{aligned} \tag{1.194}$$

where the second equality follows from the matrix relations

$$|\det[\mathbf{A}]|\det[\mathbf{K}]^{1/2} = \det[\mathbf{AKA}']^{1/2} \tag{1.195}$$

$$(\mathbf{A}^{-1}\mathbf{u}-\boldsymbol{\mu})'\mathbf{K}^{-1}(\mathbf{A}^{-1}\mathbf{u}-\boldsymbol{\mu}) = (\mathbf{u}-\mathbf{A}\boldsymbol{\mu})'(\mathbf{AKA}')^{-1}(\mathbf{u}-\mathbf{A}\boldsymbol{\mu}) \tag{1.196}$$

1.7. Laws of Large Numbers

Some of the most fundamental theorems of probability concern limiting properties for sums of random variables. This section introduces various types of convergence used in probability theory and discusses laws of large numbers and the central limit theorem. The weak law of large numbers is proven and the strong law is stated without proof. Three forms of the central limit theorem for sequences of independent random variables are discussed, all without proof. We state the form for identically distributed random

variables that is typically stated in statistics books, and also discuss Liapounov's and Lindberg's conditions for the central limit theorem to apply to independent random variables that are not necessarily identically distributed. While these may be outside the ordinary sphere of application, when interpreted for sequences of uniformly bounded random variables they help to explain the naturalness of the central limit theorem.

1.7.1. Weak Law of Large Numbers

Averages play an important role in both probability and statistics. From an empirical perspective, if a random variable X is observed n times (meaning rigorously that n identically distributed random variables are observed) and the numerical average of the observations is taken, we would like to quantify the degree to which that average can be taken as an estimate of the mean of X. More generally, what is the relationship between an average of random variables and the average of their means?

Let $X_1, X_2,...$ be random variables defined on a common sample space and possessing finite second moments. Let their means be $\mu_1, \mu_2,...$, respectively, and let

$$Y_n = \frac{1}{n}\sum_{k=1}^{n} X_k \tag{1.197}$$

be the arithmetic mean of $X_1, X_2,..., X_n$. Owing to the linearity of expectation,

$$E[Y_n] = \frac{1}{n}\sum_{k=1}^{n} \mu_k \tag{1.198}$$

and, according to Theorem 1.15,

$$\mathrm{Var}[Y_n] = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{Cov}[X_j, X_k] \tag{1.199}$$

For $\varepsilon > 0$, applying Chebyshev's inequality to Y_n yields

$$P(|Y_n - E[Y_n]| \geq \varepsilon) \leq \frac{1}{\varepsilon^2 n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{Cov}[X_j, X_k] \tag{1.200}$$

The probability on the left will converge to 0 as $n \to \infty$ if the sum of the covariances divided by n^2 converges to 0 as $n \to \infty$. This observation gives rise to Markov's form of the *weak law of large numbers*.

Theorem 1.16. Suppose X_1, X_2,... are random variables defined on a common sample space, possessing finite second moments, and having means μ_1, μ_2,..., respectively. If

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \mathrm{Cov}[X_j, X_k] = 0 \tag{1.201}$$

then, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left| \frac{1}{n} \sum_{k=1}^{n} X_k - \frac{1}{n} \sum_{k=1}^{n} \mu_k \right| \geq \varepsilon \right) = 0 \quad \blacksquare \tag{1.202}$$

The theorem asserts that, given an arbitrarily small quantity $\varepsilon > 0$, the probability that the difference between the average of the random variables and the average of their means exceeds ε can be made arbitrarily small by averaging a sufficient number of random variables.

The weak law of large numbers represents a type of convergence involving a probability measure. In general, a sequence of random variables Z_1, Z_2,... is said to *converge in probability* to random variable Z if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(|Z_n - Z| \geq \varepsilon \right) = 0 \tag{1.203}$$

The weak law of large numbers asserts that the difference between the average of a sequence of random variables and the average of their means converges to 0 in probability if the limit of Eq. 1.201 holds.

The weak law of large numbers is usually employed for specific cases. One particular case occurs when X_1, X_2,... are uncorrelated. If σ_1^2, σ_2^2,... are the variances of X_1, X_2,..., respectively, and there exists a bound M such that $\sigma_k^2 \leq M$ for $k = 1, 2,...$, then

$$\frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \mathrm{Cov}[X_j, X_k] = \frac{1}{n^2} \sum_{k=1}^{n} \sigma_k^2 \leq \frac{M}{n} \tag{1.204}$$

Consequently, the limit of Eq. 1.201 is satisfied and the weak law holds for X_1, X_2,.... If X_1, X_2,... are uncorrelated and there exists a bound M such that $\sigma_k^2 \leq M$, and, moreover, X_1, X_2,... possess a common mean μ, then the weak law states that

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right| \geq \varepsilon\right) = 0 \tag{1.205}$$

This is its most commonly employed form. If $X_1, X_2,\ldots$ are independent and identically distributed with common mean μ and finite variance σ^2, then they are uncorrelated, Eq. 1.204 holds with $M = \sigma^2$, and, as stated in Eq. 1.205, the arithmetic mean of the random variables converges in probability to their common mean.

The type of convergence that will play the dominant role in our study of random processes is defined via the second moments of the differences between the random variables of the sequence and the limiting random variable. The sequence of random variables $Z_1, Z_2,\ldots$ is said to *converge in the mean-square* to random variable Z if

$$\lim_{n\to\infty} E[|Z_n - Z|^2] = 0 \tag{1.206}$$

It is an immediate consequence of Chebyshev's inequality that, if Z_n converges to Z in the mean-square, then Z_n converges to Z in probability. The converse, however, is not valid: it is possible for a sequence of random variables to converge in probability but not in the mean-square.

***Example* 1.25.** Consider an infinite sequence of independent trials and assume that a certain event A occurs with probability p_n on trial n. If the random variable X_n is defined by $X_n = 1$ if event A occurs on trial n and $X_n = 0$ if event A does not occur on trial n, then X_n is a binomial random variable with parameters 1 and p_n. Therefore, $E[X_n] = p_n$ and $\text{Var}[X_n] = p_n(1 - p_n)$. In this context, the random variable Y_n of Eq. 1.197 is the relative frequency of event A on the first n trials. Since the random variables constituting the sequence are uncorrelated and $\text{Var}[X_n] \leq 1$ for all n, the weak law of large numbers applies and the relative frequency converges in probability to the arithmetic mean of the trial probabilities of event A. This proposition was proved by Poisson. Now, if the trial probabilities are constant, $p_n = p$, then for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|Y_n - p\right| \geq \varepsilon\right) = 0$$

and the relative frequency converges in probability to the common probability of event A. This simplified result was proved by Jacob Bernoulli and represents the first law of large numbers in probability theory. ■

1.7.2. Strong Law of Large Numbers

A random variable is a real-valued function defined on a sample space. Therefore convergence of a sequence of random variables can be considered from the perspective of ordinary function convergence. If $(S, \mathcal{E}, P)$ is a probability space, Z_n: $S \to \Re$ is a random variable for $n = 1, 2,....$, and Z: $S \to \Re$, then, as functions, Z_n converges to Z if, for any $w \in S$,

$$\lim_{n\to\infty} Z_n(w) = Z(w) \tag{1.207}$$

If two random variables differ only on a set of probability zero, then they are identically distributed. Hence, it is sufficient to have the limit of Eq. 1.207 hold everywhere in S except for an event of probability zero. We make the following definition: Z_n *converges almost surely* to Z if there exists an event G such that $P(G) = 0$ and the limit of Eq. 1.207 holds for all $w \in S - G$. This means that, for any $w \in S - G$ and any $\varepsilon > 0$, there exists a positive integer $N_{w,\varepsilon}$ such that, for $n \geq N_{w,\varepsilon}$,

$$|Z_n(w) - Z(w)| < \varepsilon \tag{1.208}$$

Almost-sure convergence is often expressed by

$$P\left(\lim_{n\to\infty} Z_n = Z\right) = 1 \tag{1.209}$$

To investigate the relationship between convergence in probability and almost-sure convergence, we introduce another type of convergence, which generalizes the classical concept of uniform convergence for functions: Z_n *converges almost uniformly* to Z if, for any $\delta > 0$, there exists an event F_δ such that $P(F_\delta) < \delta$ and Z_n converges uniformly to Z on $S - F_\delta$. Uniform convergence on $S - F_\delta$ means that, for any $\varepsilon > 0$, there exists a positive integer $N_{\delta,\varepsilon}$ such that, for $n \geq N_{\delta,\varepsilon}$, Eq. 1.208 holds for all $w \in S - F_\delta$. It is a fundamental property (*Egoroff's theorem*) of random variables that almost-sure and almost-uniform convergence are equivalent (under the assumption that we restrict our attention to random variables that are finite, except perhaps on sets of zero probability). Consequently, to show that almost-sure convergence implies convergence in probability, we need only show that almost-uniform convergence implies convergence in probability. To do so, consider arbitrary $\delta > 0$ and $\varepsilon > 0$, and choose F_δ and $N_{\delta,\varepsilon}$ so that Eq. 1.208 holds for $n \geq N_{\delta,\varepsilon}$ and $w \in S - F_\delta$. Then, for $n \geq N_{\delta,\varepsilon}$,

$$\begin{aligned} P(|Z_n - Z| \geq \varepsilon) &= P(\{w \in S: |Z_n(w) - Z(w)| \geq \varepsilon\}) \\ &= 1 - P(\{w \in S: |Z_n(w) - Z(w)| < \varepsilon\}) \end{aligned} \tag{1.210}$$

$$\leq 1 - P(S - F_\delta)$$

$$= P(F_\delta)$$

which is less than δ. Since δ has been chosen arbitrarily, the limit of Eq. 1.203 holds for arbitrary ε and Z_n converges to Z in probability. The converse is not true: convergence in probability does not imply almost-sure convergence. Consequently, almost-sure convergence is a stronger form of convergence than convergence in probability. We now state Kolmogorov's *strong law of large numbers*.

Theorem 1.17. If $X_1, X_2, \ldots$ are independent and identically distributed random variables possessing finite mean μ, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \mu \quad \text{(almost surely)} \qquad \blacksquare \qquad (1.211)$$

The strong law is equivalently written as

$$P\left(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \mu \right) = 1 \qquad (1.212)$$

The conclusion of Theorem 1.17 is stronger than the conclusion of Theorem 1.16, but so too is the hypothesis.

1.7.3. Central Limit Theorem

Inclusion probabilities of a random variable are determined by its probability distribution function (or density). If we are interested in approximating probabilities for a random variable X from probabilities of a sequence of random variables $X_1, X_2, \ldots$ converging (in some sense) to X, then we need to be concerned about the relationship between the probability distribution functions of $X_1, X_2, \ldots$ and the probability distribution function of X. X_n is said to *converge in law* (*converge in distribution*) to X if, for any point a at which the probability distribution function F_X of X is continuous,

$$\lim_{n\to\infty} F_{X_n}(a) = F_X(a) \qquad (1.213)$$

Equivalently, if F_X is continuous at a and b, $a < b$, then

$$\lim_{n\to\infty}\left(F_{X_n}(b)-F_{X_n}(a)\right)=F_X(b)-F_X(a) \tag{1.214}$$

which, in terms of interval probabilities, means that

$$\lim_{n\to\infty} P(a<X_n\le b)=P(a<X\le b) \tag{1.215}$$

If X is a continuous random variable, then its probability distribution function is continuous and the preceding limit can be expressed in terms of densities as

$$\lim_{n\to\infty}\int_a^b f_{X_n}(x)\,dx=\int_a^b f_X(x)\,dx \tag{1.216}$$

Relative to our original point concerning approximation, if X_n converges to X in law, then, for large n,

$$P(a<X_n\le b)\approx P(a<X\le b) \tag{1.217}$$

and probabilities of X can be used to approximate probabilities of X_n.

As with laws of large numbers, our concern is with limits of averages, or sums, of sequences of independent random variables. Here, however, our interest lies in the limiting distributions of the averages. The manner in which distributions of averages converge to the standard normal distribution (that is, how averages of nonnormal distributions converge in law to the standard normal distribution) explains the special role of the normal distribution in probability and statistics, a role recognized by Gauss. We first state the most commonly employed form of the *central limit theorem* and subsequently discuss more general formulations. The first form of the theorem is stated in terms of standardized random variables.

For a random variable X with mean μ and variance σ^2, define the *standardized* random variable $(X-\mu)/\sigma$. The standardized variable has zero mean and unit variance. If $X_1, X_2,\ldots$ are independent, identically distributed random variables possessing common mean μ and variance σ^2, then the average Y_n of Eq. 1.197 has mean μ and variance σ^2/n. Hence, the standardized variable corresponding to Y_n is $(Y_n-\mu)\sqrt{n}/\sigma$. The central limit theorem asserts that this standardized average converges in law to the standard normal random variable.

Theorem 1.18. If $X_1, X_2,\ldots$ are independent, identically distributed random variables possessing common mean μ and variance σ^2, then, for any z,

$$\lim_{n\to\infty} P\left(\frac{\frac{1}{n}\sum_{k=1}^{n} X_k - \mu}{\frac{\sigma}{\sqrt{n}}} \le z \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-y^2/2}\, dy \qquad \blacksquare \qquad (1.218)$$

If we denote the standardized average by

$$Z_n = \frac{\frac{1}{n}\sum_{k=1}^{n} X_k - \mu}{\sigma/\sqrt{n}} \qquad (1.219)$$

the standard normal random variable by Z, and the probability distribution function of the standard normal variable by $\Phi(z)$, then the central limit theorem can be written simply as

$$\lim_{n\to\infty} Z_n = Z \qquad \text{(in law)} \qquad (1.220)$$

or, in terms of interval probabilities, as

$$\lim_{n\to\infty} P(a < Z_n \le b) = \Phi(b) - \Phi(a) \qquad (1.221)$$

***Example* 1.26.** The individual trials of the binomial distribution with success probability p are independent and identically distributed. Let $X_k = 1$ if there is a success on trial k and $X_k = 0$ if there is a failure on trial k. The binomial random variable for n trials is given by

$$X^n = \sum_{k=1}^{n} X_k$$

$X_1, X_2, \ldots$ possess common mean p and variance $p(1-p)$, X^n has mean np and variance $np(1-p)$, the average X^n/n has mean p and variance $p(1-p)/n$, and the central limit theorem states that

$$\lim_{n\to\infty} \frac{X^n/n - p}{\sqrt{p(1-p)/n}} = Z \qquad \text{(in law)}$$

In terms of interval probabilities,

$$\lim_{n\to\infty} P\left(a < \frac{X^n - np}{\sqrt{np(1-p)}} \le b \right) = \Phi(b) - \Phi(a)$$

This is the classical *DeMoivre-Laplace* theorem. It can be used to estimate binomial probabilities via the standard normal distribution. Writing the preceding limit as an approximation for large n and changing variables yields

$$P(a < X^n \le b) \approx \Phi\left(\frac{b - np}{\sqrt{np(1-p)}} \right) - \Phi\left(\frac{a - np}{\sqrt{np(1-p)}} \right) \qquad \blacksquare$$

There are more general conditions under which the central limit theorem applies to a sequence of independent random variables $X_1, X_2, \ldots$ that need not be identically distributed. The most celebrated of these, which we state next, is due to Liapounov and for this reason the central limit theorem is often called *Liapounov's theorem*. To simplify expressions, we employ the following notation for the remainder of the section, under the assumption that $X_1, X_2, \ldots$ are independent random variables with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$, respectively. Let

$$S_n = \sum_{k=1}^{n} X_k \tag{1.222}$$

$$m_n = \sum_{k=1}^{n} \mu_k \tag{1.223}$$

$$s_n^2 = \sum_{k=1}^{n} \sigma_k^2 = \sum_{k=1}^{n} E[|X_k - \mu_k|^2] \tag{1.224}$$

S_n is the sum of the first n random variables in the sequence, m_n is the mean of S_n, and, because the random variables are independent, s_n^2 is the variance of S_n.

Theorem 1.19. Suppose $X_1, X_2, \ldots$ are independent random variables with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$, respectively. If there exists $\delta > 0$ such that $E[|X_n - \mu_n|^{2+\delta}]$ is finite for $n = 1, 2, \ldots$ and

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} E[|X_k - \mu_k|^{2+\delta}] = 0 \tag{1.225}$$

then

$$\lim_{n\to\infty} \frac{S_n - m_n}{s_n} = Z \quad \text{(in law)} \qquad \blacksquare \tag{1.226}$$

The limiting condition of Eq. 1.225, known as *Liapounov's condition*, represents the profound part of the theorem. It is not easily apprehended; however, from a practical perspective, the theorem has an easily applied corollary: if there exists a fixed bound C such that $|X_n| \leq C$ for all n and

$$\lim_{n\to\infty} s_n^2 = \sum_{k=1}^{\infty} \sigma_k^2 = \infty \tag{1.227}$$

then the limiting conclusion of the central limit theorem applies. In fact, the uniform boundedness of the random variables implies

$$\begin{aligned} E[|X_k - \mu_k|^{2+\delta}] &= E[|X_k - \mu_k|^{\delta} |X_k - \mu_k|^2] \\ &\leq 2^{\delta} C^{\delta} E[|X_k - \mu_k|^2] \end{aligned} \tag{1.228}$$

Hence,

$$\frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} E[|X_k - \mu_k|^{2+\delta}] \leq \frac{2^{\delta} C^{\delta}}{s_n^{\delta} s_n^2} \sum_{k=1}^{n} E[|X_k - \mu_k|^2] = \frac{2^{\delta} C^{\delta}}{s_n^{\delta}} \tag{1.229}$$

which, according to the assumption of Eq. 1.227, converges to 0 as $n \to \infty$. While we might model random variables by unbounded distributions, real-world data are uniformly bounded. Moreover, divergence of the sum of the variances can be expected from real-world random phenomena since, for the sum of the variances to converge, it would be necessary that $\mathrm{Var}[X_n] \to 0$ as $n \to \infty$. Consequently, convergence of the distribution of $(S_n - m_n)/s_n$ to the standard normal distribution can be expected for many natural phenomena.

There is a more general sufficient condition than that of Liapounov for the central limit theorem. *Lindberg's condition* states that the central limit theorem (Eq. 1.226) holds for a sequence of independent random variables if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} \int_{\{x:|x-\mu_k|>\varepsilon s_n\}} (x-\mu_k)^2 f_k(x)\, dx = 0 \tag{1.230}$$

where $f_k(x)$ is the density for X_k. Not only is Lindberg's condition sufficient, with a slight modification in the statement of the theorem, it is also necessary. Specifically, we have the following restatement of the central limit theorem: for a sequence of independent random variables, there is convergence in law according to Eq. 1.226 and

$$\lim_{n\to\infty} \frac{\max_{k\le n} \sigma_k}{s_n} = 0 \tag{1.231}$$

if and only if Lindberg's condition holds. The condition of Eq. 1.231, which when added to convergence in law of $(S_n - m_n)/s_n$ to Z, makes Lindberg's condition necessary and sufficient, is due to Feller and states that the maximum variance among the random variables in the sequence up to n becomes negligible relative to the variance of the sum as $n \to \infty$.

In conclusion, one should recognize the long history of convergence theorems in the theory of probability. The great scientific names associated with the various theorems and the degree to which they have remained a central theme are testimony to their importance. The central limit theorem alone shows the profound relationship between theoretical mathematics, applied science, and the philosophy of nature.

1.8. Parametric Estimation via Random Samples

If we model a random variable by some probability distribution (normal, gamma, etc.), there remains the task of specifying the distributional parameters in such a way that the probability mass reflects the distribution of the phenomena described by the random variable. For instance, suppose X is assumed to be gamma distributed. Then X has a density $f(x; \alpha, \beta)$, where the notation is meant to imply that the parameters α and β are unknown and need to be particularized to X. From observations of random variables related to X, we wish to estimate the relevant values of the parameters.

1.8.1. Random-Sample Estimators

A common way of proceeding is to observe a set of random variables X_1, $X_2,\ldots, X_n$ that are both independent and identically distributed to X. Such a set of random variables is called a *random sample* for X. If X has density $f(x)$, then, owing to independence and identical distribution, the joint density of $X_1, X_2,\ldots, X_n$ is given by

$$f(x_1, x_2,\ldots, x_n) = \prod_{k=1}^{n} f(x_k) \tag{1.232}$$

Now suppose X has density $f(x; \theta)$ with unknown parameter θ. A function of a random sample $X_1, X_2, \ldots, X_n$, say,

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n) \tag{1.233}$$

is needed to provide estimation of θ based on $X_1, X_2, \ldots, X_n$. $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is itself a random variable and is called an *estimator* of θ. Given sample values $x_1, x_2, \ldots, x_n$ from observations of $X_1, X_2, \ldots, X_n$, the functional value $\hat{\theta}(x_1, x_2, \ldots, x_n)$ provides an *estimate* of θ for the particular sample. The function rule defining $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is called the *estimation rule* for the estimator. $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is called a *statistic* if the estimation rule defining it is free of unknown parameters (which does not mean that the distribution of $\hat{\theta}$ is free of unknown parameters). We assume that estimators are statistics.

Two sets of sample values, $\{x_1, x_2, \ldots, x_n\}$ and $\{z_1, z_2, \ldots, z_n\}$, obtained from observation of a random sample $X_1, X_2, \ldots, X_n$ give rise to two, almost certainly distinct, estimates of θ, $\hat{\theta}(x_1, x_2, \ldots, x_n)$ and $\hat{\theta}(z_1, z_2, \ldots, z_n)$. Is one of the estimates better than the other? Is either a good estimate? As posed, these questions are not meaningful. Goodness must be posed in terms of the estimator (estimation rule). Whether an estimator provides "good" estimates depends on positing some measure of goodness and then examining the estimator relative to the measure of goodness.

An estimator $\hat{\theta}$ of θ is said to be *unbiased* if $E[\hat{\theta}] = \theta$. Implicit in the definition is that $E[\hat{\theta}] = \theta$ regardless of the value of θ. $E[\hat{\theta}] - \theta$ is called the *bias* of $\hat{\theta}$. Since $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$, $\hat{\theta}$ depends on the sample size n. In many circumstances an estimator is biased but the bias diminishes as the sample size increases. Estimator $\hat{\theta}$ is said to be *asymptotically unbiased* if $E[\hat{\theta}] \to \theta$ as $n \to \infty$. Unbiasedness is a desirable property because it means that on average the estimator provides the desired parameter. However, this may be of little practical value if the variance of the estimator is too great.

To desire *precision* in an estimator is to wish the estimator to be within some tolerance of the parameter. Since the estimator is random, this cannot be guaranteed; however, for $r > 0$, we can consider the probability that the estimator is within r of the parameter, namely, $P(|\hat{\theta} - \theta| < r)$. $\hat{\theta}$ is said to be a *consistent* estimator of θ if, for any $r > 0$,

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| < r) = 1 \tag{1.234}$$

An equivalent way of stating the matter is that $\hat{\theta}$ is a consistent estimator of θ if and only if $\hat{\theta}$ converges in probability to θ.

If $\hat{\theta}$ is unbiased, then, according to Chebyshev's inequality,

$$P(|\hat{\theta}-\theta|<r)\geq 1-\frac{\mathrm{Var}[\hat{\theta}]}{r^2} \tag{1.235}$$

Consequently, if $\hat{\theta}$ is unbiased and $\mathrm{Var}[\hat{\theta}] \to 0$ as $n \to \infty$, then $\hat{\theta}$ is a consistent estimator. This proposition can be strengthened: $\hat{\theta}$ is a consistent estimator of θ if $\hat{\theta}$ is an asymptotically unbiased estimator of θ and $\mathrm{Var}[\hat{\theta}] \to 0$ as $n \to \infty$. Indeed, asymptotic unbiasedness implies that $|E[\hat{\theta}] - \theta| < r/2$ for sufficiently large n, so that

$$\begin{aligned} P(|\hat{\theta}-\theta|\geq r) &\leq P(|\hat{\theta}-E[\hat{\theta}]|+|E[\hat{\theta}]-\theta|\geq r) \\ &\leq P\left(\left(|\hat{\theta}-E[\hat{\theta}]|\geq\frac{r}{2}\right)\cup\left(|\theta-E[\hat{\theta}]|\geq\frac{r}{2}\right)\right) \\ &= P\left(|\hat{\theta}-E[\hat{\theta}]|\geq\frac{r}{2}\right) \\ &\leq \frac{4\mathrm{Var}[\hat{\theta}]}{r^2} \end{aligned} \tag{1.236}$$

for sufficiently large n (the last inequality resulting from Chebyshev's inequality).

1.8.2. Sample Mean and Sample Variance

Given a random sample $X_1, X_2, \ldots, X_n$ for the random variable X possessing mean μ, a commonly employed estimator of μ is the *sample mean*

$$\overline{X}=\frac{X_1+X_2+\cdots+X_n}{n} \tag{1.237}$$

The distribution of the sample mean (as a random variable) is called the *sampling distribution of the mean*. Generically, the sample-mean estimation rule is the function

$$\hat{\theta}(\xi_1,\xi_2,\ldots,\xi_n)=\frac{\xi_1+\xi_2+\cdots+\xi_n}{n} \tag{1.238}$$

Applied to the random sample $X_1, X_2, \ldots, X_n$, this estimation rule gives the sample-mean estimator. If sample values $x_1, x_2, \ldots, x_n$ are observed and the

sample-mean estimator is used, then the estimation rule applied to $x_1, x_2,..., x_n$ yields an estimate of μ,

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.239}$$

called an *empirical mean*.

The sample mean is unbiased, $E[\bar{X}] = \mu$, and, by Theorem 1.15, $\text{Var}[\bar{X}] = \sigma^2/n$. Since the sample mean is unbiased and its variance tends to 0 as $n \to \infty$, it is a consistent estimator of the mean. Consistency of the sample mean is a form of the weak law of large numbers for the special case of a random sample. In fact, the complementary form of Eq. 1.200 applied to the special case of an average of a random sample is given by

$$P(|\bar{X} - \mu| < r) \geq 1 - \frac{\sigma^2}{nr^2} \tag{1.240}$$

which is itself the complementary form of Chebyshev's inequality given in Eq. 1.75 as applied to the sample mean.

Not only does the sample mean converge to the mean in probability, but according to the strong law of large numbers, it converges to the mean almost surely. Furthermore, according to the central limit theorem, the standardized version of the sample mean converges in law to the standard normal random variable. Consequently, for large n,

$$P\left(a < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < b\right) \approx \Phi(b) - \Phi(a) \tag{1.241}$$

where $\Phi(z)$ is the probability distribution function for the standard normal variable. This approximation is used to compute approximate probabilities concerning the sample mean via the standard normal probability distribution.

The most common variance estimator is the *sample variance*, which for a random sample $X_1, X_2,..., X_n$ arising from the random variable X, is defined by

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2 \tag{1.242}$$

The variance is an unbiased estimator of the variance:

$$E[S^2] = \frac{1}{n-1}E\left[\sum_{k=1}^{n}(X_k - \bar{X})^2\right]$$

$$= \frac{1}{n-1} E\left[\sum_{k=1}^{n} (X_k - \mu)^2 - n(\bar{X} - \mu)^2\right]$$

$$= \frac{1}{n-1}\left(\sum_{k=1}^{n} E[(X_k - \mu)^2] - nE[(\bar{X} - \mu)^2]\right) \tag{1.243}$$

$$= \frac{1}{n-1}\left(\sum_{k=1}^{n} \mathrm{Var}[X_k] - n\mathrm{Var}[\bar{X}]\right)$$

$$= \mathrm{Var}[X]$$

where the second equality results from some algebraic manipulation and the last from the identical distribution of $X_1, X_2, \ldots, X_n$ and the fact that $\mathrm{Var}[\bar{X}] = \mathrm{Var}[X]/n$.

A basic theorem of statistics states that, for a random sample $X_1, X_2, \ldots, X_n$ arising from a normally distributed random variable X with mean μ and variance σ^2, $\bar{X}$ and S^2 are independent and $(n-1)S^2/\sigma^2$ possesses a gamma distribution with $\alpha = (n-1)/2$ and $\beta = 2$. Since the variance of the gamma distribution is $\alpha\beta^2$,

$$\mathrm{Var}[S^2] = \mathrm{Var}\left[\frac{\sigma^2}{n-1}\left(\frac{(n-1)S^2}{\sigma^2}\right)\right] = \frac{2\sigma^4}{n-1} \tag{1.244}$$

Hence, $\mathrm{Var}[S^2] \to 0$ as $n \to \infty$. Since S^2 is an unbiased estimator, for a normal random variable the sample variance is a consistent estimator of the variance. Practically, however, there is a difficulty. If the variance of X is large, then the variance of the sample variance will be large because the original variance is squared and then doubled in the numerator of $\mathrm{Var}[S^2]$. The sample size may have to be prohibitively large to obtain a sufficiently small variance of the sample variance. Specifically, the complementary form of Chebyshev's inequality takes the form

$$P(|S^2 - \sigma^2| < r) \geq 1 - \frac{2\sigma^4}{(n-1)r^2} \tag{1.245}$$

The precision of the sample variance compared with the precision of the sample mean can be seen by comparing Eqs. 1.245 and 1.240.

1.8.3. Minimum-Variance Unbiased Estimators

Application of Chebyshev's inequality in Eq. 1.235 yields a lower bound on the precision $P(|\hat{\theta} - \theta| < r)$ in terms of the variance of $\hat{\theta}$ under the assumption that $\hat{\theta}$ is an unbiased estimator. From a limiting perspective, if $\mathrm{Var}[\hat{\theta}] \to 0$ as $n \to \infty$, then $\hat{\theta}$ is a consistent estimator, meaning $\hat{\theta} \to \theta$ in probability as $n \to \infty$. From the standpoint of comparing two unbiased estimators of θ, the one with smaller variance gives a greater lower bound.

Another approach is to consider the mean-square error, $E[|\hat{\theta} - \theta|^2]$, of $\hat{\theta}$ as an estimator of θ. Instead of consistency, we can ask whether $\hat{\theta}$ converges to θ in the mean-square; that is, does

$$\lim_{n\to\infty} E[|\hat{\theta}-\theta|^2] = 0 \tag{1.246}$$

The mean-square error can be expanded in terms of the variance and bias as

$$\begin{aligned} E[|\hat{\theta} - \theta|^2] &= E[\hat{\theta}^2] - 2E[\hat{\theta}\,\theta] + E[\theta^2] \\ &= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\ &= \mathrm{Var}[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2 \end{aligned} \tag{1.247}$$

If $\hat{\theta}$ is not asymptotically unbiased, then $(E[\hat{\theta}] - \theta)^2$ does not converge to 0 and $\hat{\theta}$ cannot converge to θ in the mean-square: if $\hat{\theta}$ is asymptotically unbiased, then

$$\lim_{n\to\infty} E[|\hat{\theta}-\theta|^2] = \lim_{n\to\infty} \mathrm{Var}[\hat{\theta}] \tag{1.248}$$

and $\hat{\theta} \to \theta$ in the mean-square as $n \to \infty$ if and only if $\mathrm{Var}[\hat{\theta}] \to 0$ as $n \to \infty$. Recall that mean-square convergence implies convergence in probability (consistency).

If $\hat{\theta}$ is unbiased, then

$$E[|\hat{\theta} - \theta|^2] = \mathrm{Var}[\hat{\theta}] \tag{1.249}$$

Given unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ, the one with smaller variance has a smaller mean-square error. Hence, we call $\hat{\theta}_1$ a *better* estimator than $\hat{\theta}_2$ if $\mathrm{Var}[\hat{\theta}_1] < \mathrm{Var}[\hat{\theta}_2]$, where it is implicit that the estimators are being compared for the same sample size. More generally, an unbiased estimator $\hat{\theta}$ is said to be a *minimum-variance unbiased estimator* (*MVUE*) of θ if for any other unbiased estimator $\hat{\theta}_0$ of θ, $\mathrm{Var}[\hat{\theta}] \le \mathrm{Var}[\hat{\theta}_0]$. $\hat{\theta}$ is also called a *best*

unbiased estimator of θ. Even if an MVUE cannot be found, it may be possible to find an unbiased estimator of θ that has minimum variance among all estimators of θ in some restricted class C of estimators. Such an estimator is a best estimator relative to C.

***Example* 1.27.** Let $X_1, X_2, \ldots, X_n$ be a random sample from a random variable X with mean μ. Let C be the class of all linear unbiased estimators of μ, these being of the form

$$\hat{\mu} = \sum_{k=1}^{n} a_k X_k$$

with $E[\hat{\mu}] = \mu$. By linearity of the expectation,

$$E[\hat{\mu}] = \left(\sum_{k=1}^{n} a_k \right) \mu$$

Therefore it must be that $a_1 + a_2 + \cdots + a_n = 1$. The variance of $\hat{\mu}$ is

$$\sigma_{\hat{\mu}}^2 = \sum_{k=1}^{n} a_k^2 \sigma^2 = \left(\sum_{k=1}^{n-1} a_k^2 + \left(1 - \sum_{k=1}^{n-1} a_k \right)^2 \right) \sigma^2$$

For $j = 1, 2, \ldots, n-1$,

$$\frac{\partial}{\partial a_j} \sigma_{\hat{\mu}}^2 = \left(a_j - 1 + \sum_{k=1}^{n-1} a_k \right) 2\sigma^2$$

Setting the derivative equal to 0 yields

$$a_1 + \cdots + a_{j-1} + 2a_j + a_{j+1} + \cdots + a_{n-1} = 1$$

Setting all derivatives, $j = 1, 2, \ldots, n-1$, to 0 yields the system

$$\begin{pmatrix} 2 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ 1 & 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

The system is solved by $a_j = 1/n$ for $j = 1, 2,..., n - 1$. Since the system matrix is nonsingular, this solution is unique. Moreover, since the sum of the coefficients is 1, $a_n = 1/n$ and the best linear unbiased estimator is the sample mean. ■

Finding a minimum-variance unbiased estimator is generally difficult; however, checking whether a particular unbiased estimator has minimum variance can often be accomplished with the aid of the next theorem, known as the *Cramer-Rao inequality*. If an estimator has a variance equal to the lower bound stated in the theorem, then the estimator must be an MVUE. There are a number of regularity conditions on the density whose parameter is being estimated. These have to do with existence of the partial derivative in the Cramer-Rao lower bound, finiteness and positivity of the expectation in the bound denominator, and sufficient regularity of the density to bring partial derivatives inside expectation-defining integrals in the proof of the theorem. We omit them from the statement of the theorem, noting that they are satisfied for commonly employed densities.

Theorem 1.20. Under certain regularity conditions, if $X_1, X_2,..., X_n$ comprise a random sample arising from a random variable X and $\hat{\theta}$ is an unbiased estimator of the parameter θ in the density $f(x; \theta)$ of X, then

$$\text{Var}[\hat{\theta}] \geq \frac{1}{nE\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right]} \quad \blacksquare \qquad (1.250)$$

***Example* 1.28.** We demonstrate that the sample mean is an MVUE for the mean of a normally distributed random variable. If X has mean μ and variance σ^2, then

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\log f(x;\mu) = -\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

$$\frac{\partial}{\partial\mu}\log f(x;\mu) = \frac{x-\mu}{\sigma^2}$$

$$E\left[\left(\frac{\partial}{\partial\mu}\log f(X;\mu)\right)^2\right] = \frac{E[(X-\mu)^2]}{\sigma^4} = \frac{1}{\sigma^2}$$

Hence, the Cramer-Rao lower bound is σ^2/n, the variance of the sample mean. ■

There is crucial difference between the results of Examples 1.27 and 1.28. The former shows that the sample mean is the best estimator of the mean among the class of linear unbiased estimators, regardless of the distribution; the latter shows that the sample mean is the best estimator among all unbiased estimators of the mean of a normal distribution. In fact, the sample mean is not always the MVUE for the mean of a distribution.

1.8.4. Method of Moments

Just because an estimator is best for one distribution does not imply it is best for another distribution. In many cases one cannot find a best estimator and therefore settles for one that performs satisfactorily. There are a number of techniques for finding estimation rules. The same technique applied to different distributions often produces different estimation rules. Bestness depends on both the distribution and the measure of goodness. More interesting, different techniques can yield different estimators for the same distribution. In addition, the properties of the estimator produced by a technique for finding estimation rules will vary, depending on the distribution to which it is applied. Two commonly used techniques for finding estimators are maximum-likelihood and the method of moments. Generally, the properties of maximum-likelihood estimators are preferable to method-of-moment estimators, but the method of moments may be applied in many circumstances where maximum likelihood is mathematically intractable. Here we consider the method of moments; the next section is devoted to maximum-likelihood estimation.

If $X_1, X_2,\ldots, X_n$ comprise a random sample arising from the random variable X, then the rth *sample moment* of $X_1, X_2,\ldots, X_n$ is

$$M_r' = \frac{1}{n}\sum_{k=1}^{n} X_k^r \tag{1.251}$$

For $r = 1$, the sample moment is the sample mean. For $r = 2$,

$$M_2' = \frac{n-1}{n}S^2 + \overline{X}^2 \tag{1.252}$$

For $r = 1, 2,...,$ M_r' is the arithmetic mean of $X_1^r, X_2^r, \ldots, X_n^r$, which themselves constitute a random sample arising from the random variable X^r. Consequently, if $E[X^r]$, the rth moment about the origin of X, exists, then M_r' is an unbiased estimator of $E[X^r]$, so that

$$E[M_r'] = E[X^r] \tag{1.253}$$

$$\text{Var}[M_r'] = \frac{\text{Var}[X^r]}{n} \tag{1.254}$$

and M_r' is a consistent estimator of $E[X^r]$.

If X has density $f(x; \theta_1, \theta_2,..., \theta_p)$, where parameters $\theta_1, \theta_2,..., \theta_p$ need to be estimated, then $E[X^r]$ is a function of $\theta_1, \theta_2,..., \theta_p$, meaning there exists a function h_r such that

$$E[X^r] = h_r(\theta_1, \theta_2, \ldots, \theta_p) \tag{1.255}$$

Method-of-moments estimation is done by setting $E[X^r] = M_r'$, which is reasonable since M_r' is a consistent estimator of $E[X^r]$. This leads to the system of equations

$$\begin{aligned} h_1(\theta_1, \theta_2, \ldots, \theta_p) &= M_1' \\ h_2(\theta_1, \theta_2, \ldots, \theta_p) &= M_2' \\ &\vdots \\ h_N(\theta_1, \theta_2, \ldots, \theta_p) &= M_N' \end{aligned} \tag{1.256}$$

where N, the number of moments employed, is chosen so that a unique solution for $\theta_1, \theta_2,..., \theta_p$ can be found. The solutions $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p$ are the *method-of-moments estimators* for $\theta_1, \theta_2,..., \theta_p$. Properties of the estimators depend on the distribution and often very little is known about method-of-moment estimators. Once found, they are usually tested on data (often synthetic) to see how they perform.

***Example* 1.29.** Consider the gamma distribution with parameters α and β. Since

$$E[X] = \alpha\beta$$

$$E[X^2] = (\alpha + 1)\alpha\beta^2$$

the method-of-moments system is

$$\alpha\beta = M_1' = \overline{X}$$

$$(\alpha+1)\alpha\beta^2 = M_2' = \frac{n-1}{n}S^2 + \overline{X}^2$$

Solving for α and β yields

$$\hat{\alpha} = \frac{\overline{X}^2}{\frac{n-1}{n}S^2}$$

$$\hat{\beta} = \frac{\frac{n-1}{n}S^2}{\overline{X}}$$ ■

The system of Eq. 1.256 reveals that the key to the method is finding estimators of functions of the unknown parameters and then setting the estimators equal to the functions (as if the estimators had zero variance). As demonstrated, the method employs sample moments; however, the functions and estimators can come from other sources. In image processing the functions can provide geometric characteristics in terms of the unknown parameters and the estimators can be for these geometric characteristics.

1.8.5. Order Statistics

Sample values are often ordered from least to greatest, with some value in the ordering being of interest. If $X_1, X_2, \ldots, X_n$ comprise a random sample arising from the random variable X, then the *n order statistics* for the sample are the n random variables $Y_1 \le Y_2 \le \ldots \le Y_n$ resulting from ordering the values of the sample variables from lowest to highest. Each order statistic can be expressed as a function of the sample variables. For instance,

$$Y_1 = \min\{X_1, X_2, \ldots, X_n\} \tag{1.257}$$

$$Y_n = \max\{X_1, X_2, \ldots, X_n\} \tag{1.258}$$

A number of statistics are defined in terms of the order statistics of a random sample. Perhaps the most important is the *sample median*, which for odd n is defined by $\tilde{X} = Y_{(n+1)/2}$, the middle value of the observations. The sample median is often used to estimate the mean of symmetric distributions, especially when the distribution of the underlying random variable is heavy-

tailed, meaning that values far from the mean possess relatively high probabilities, say in comparison to the tails of a normal distribution.

Theorem 1.21. If $X_1, X_2,..., X_n$ constitute a random sample arising from the continuous random variable X and $Y_1 \le Y_2 \le ... \le Y_n$ are the n order statistics resulting from the sample, then, for $k = 1, 2,..., n$, the density of the kth order statistic is

$$f_{Y_k}(y) = \frac{n!}{(n-k)!(k-1)!} F_X(y)^{k-1}[1-F_X(y)]^{n-k} f_X(y) \tag{1.259}$$

In particular,

$$f_{Y_1}(y) = n[1-F_X(y)]^{n-1} f_X(y) \tag{1.260}$$

$$f_{Y_n}(y) = nF_X(y)^{n-1} f_X(y) \tag{1.261}$$

and, if n is odd, the density of the sample median is

$$f_{\tilde{X}}(y) = \frac{n!}{[((n-1)/2)!]^2} F_X(y)^{(n-1)/2}[1-F_X(y)]^{(n-1)/2} f_X(y) \qquad \blacksquare \tag{1.262}$$

We demonstrate the cases for the minimum and maximum order statistics, Eqs. 1.260 and 1.261, respectively. The minimum Y_1 has probability distribution function

$$\begin{aligned} F_{Y_1}(y) &= P(\min\{X_1, X_2, \ldots, X_n\} \le y) \\ &= P\left(\bigcup_{k=1}^{n} (X_k \le y)\right) \\ &= 1 - P\left(\bigcap_{k=1}^{n} (X_k > y)\right) \\ &= 1 - P(X > y)^n \\ &= 1 - (1 - F_X(y))^n \end{aligned} \tag{1.263}$$

Differentiation with respect to y gives Eq. 1.260. For the maximum Y_n,

$$F_{Y_n}(y) = P(\max\{X_1, X_2, \ldots, X_n\} \leq y)$$

$$= P\left(\bigcap_{k=1}^{n}(X_k \leq y)\right) \tag{1.264}$$

$$= F_X(y)^n$$

Differentiation with respect to y gives Eq. 1.261.

1.9. Maximum-Likelihood Estimation

Estimator properties are distribution dependent; nevertheless, maximum-likelihood estimators tend to perform quite well for commonly encountered distributions. A number of standard image/signal processing operators arise as maximum-likelihood estimators, in particular, the mean, the median, weighted medians, and flat morphological filters.

1.9.1. Maximum-Likelihood Estimators

Suppose X is a random variable with density $f(x; \theta)$, θ is a parameter to be estimated, and $X_1, X_2, \ldots, X_n$ are independent random variables identically distributed to X, so that $X_1, X_2, \ldots, X_n$ compose a random sample for X. The joint density of $X_1, X_2, \ldots, X_n$ is called the *likelihood function* of the random sample $X_1, X_2, \ldots, X_n$. Owing to independence and identical distribution with X, the likelihood function is given by

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta) \tag{1.265}$$

To simplify notation, we may write the likelihood function as $L(\theta)$.

If a value of θ can be found to maximize the likelihood function for a set of sample outcomes $x_1, x_2, \ldots, x_n$, then that value is called the *maximum-likelihood estimate* with respect to $x_1, x_2, \ldots, x_n$. If the same functional relationship between the estimate and $x_1, x_2, \ldots, x_n$ holds for all possible choices of $x_1, x_2, \ldots, x_n$, then the functional relationship is taken as an estimation rule and provides an estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ called the *maximum-likelihood estimator* (*filter*) for θ.

For intuitive appreciation, suppose X is discrete. Then

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{k=1}^{n} P(X = x_k; \theta) \tag{1.266}$$

Suppose a set of sample values $x_1, x_2,..., x_n$ is observed. Given these n observations, what value of θ would be a reasonable choice? If there exists a value θ' such that

$$L(x_1, x_2,..., x_n; \theta') \geq L(x_1, x_2,..., x_n; \theta) \tag{1.267}$$

for all θ, then θ' maximizes $P(X_1 = x_1, X_2 = x_2,..., X_n = x_n)$, which is the probability of obtaining the observations that were, in fact, obtained.

If the density of X has more than a single unknown parameter, say $f(x; \theta_1, \theta_2,..., \theta_m)$, then the definitions of likelihood function and maximum-likelihood estimator remain the same, with the vector $\boldsymbol{\theta} = (\theta_1, \theta_2,..., \theta_m)'$ taking the place of the single parameter θ. The maximum-likelihood estimator is now a vector estimator

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, X_2,..., X_n) = \begin{pmatrix} \hat{\theta}_1(X_1, X_2,..., X_n) \\ \hat{\theta}_2(X_1, X_2,..., X_n) \\ \vdots \\ \hat{\theta}_m(X_1, X_2,..., X_n) \end{pmatrix} \tag{1.268}$$

Maximum-likelihood estimators possess the *invariance* property: if $\hat{\theta}$ is a maximum-likelihood estimator of θ, g is a one-to-one function, and $\phi = g(\theta)$, then $\hat{\phi} = g(\hat{\theta})$ is a maximum-likelihood estimator of ϕ. In many cases maximum-likelihood estimators possess desirable properties, especially for large samples. For instance, the maximum- likelihood estimator is often (but not always) a minimum-variance-unbiased estimator. Indeed, there are instances in which the maximum-likelihood estimator is quite poor; however, for many commonly applied distributions, the maximum-likelihood estimator is either best or close to best. When the theoretical properties of the estimator are not known, it is prudent to perform simulations and estimate the bias and variance of the maximum-likelihood estimator for processes of interest.

***Example* 1.30.** For X normally distributed with unknown mean μ and known variance σ^2, the likelihood function is

$$L(x_1, x_2,..., x_n; \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right]$$

$L(x_1, x_2,..., x_n; \mu)$ is maximized if and only if its logarithm,

$$\log L(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{k=1}^{n}\left(\frac{x_k - \mu}{\sigma}\right)^2$$

is maximized. Differentiating with respect to μ gives

$$\frac{d}{d\mu}\log L(\mu) = \frac{1}{\sigma^2}\sum_{k=1}^{n}(x_k - \mu)$$

Setting the derivative equal to 0 shows the maximum to be the mean of x_1, $x_2,\ldots, x_n$. Since this estimation rule holds for any choice of x_1, $x_2,\ldots, x_n$, it provides the maximum-likelihood estimator $\hat{\mu} = \overline{X}$, the sample mean of the observations.

Continue to assume that X is normally distributed with mean μ and variance σ^2, but now assume that the variance is also unknown. By treating $(\mu, \sigma^2)'$ as a parameter vector, maximum-likelihood estimation can be used to find estimators of both μ and σ^2. The likelihood function remains the same, except now it is a function $L(\mu, \sigma^2)$ of both μ and σ^2. To maximize it, take its partial derivatives with respect to μ and σ^2 to obtain

$$\frac{\partial}{\partial\mu}\log L(\mu) = \frac{1}{\sigma^2}\sum_{k=1}^{n}(x_k - \mu)$$

$$\frac{\partial}{\partial\sigma^2}\log L(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{k=1}^{n}(x_k - \mu)^2$$

Simultaneous solution yields the maximum-likelihood estimators $\hat{\mu} = \overline{X}$ and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2 = \frac{n-1}{n}S^2$$

where S^2 is the sample variance. Maximum-likelihood estimation has produced an unbiased estimator of the mean and an asymptotically unbiased estimator of the variance. If the mean μ is known, then maximum-likelihood estimation applied to σ^2 alone yields the unbiased variance estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2$$

■

1.9.2. Additive Noise

A substantial portion of the text concerns estimation of signals from noise-corrupted observations. A very simple model involving additive noise has a constant (pure) discrete signal corrupted by independent, identically distributed, zero-mean additive noise. The observed signal then takes the form

$$X(i) = \theta + N(i) \tag{1.269}$$

$i = 1, 2,...$, where θ is a constant to be estimated, the noise random variables $N(i)$ possess a common distribution with variance σ^2, and any finite collection of noise random variables is independent. The observed signal $X(i)$ has mean θ and variance σ^2. We use maximum-likelihood estimation for θ, employing a finite number of observations, which for notational consistency we denote by $X_1, X_2,..., X_n$. From this perspective the maximum-likelihood estimator is a filter operating on observed signal values to estimate the pure signal θ. For instance, relative to Example 1.30, the noise is normally distributed at each point with mean 0 and variance σ^2, the observed signal values are normally distributed with mean θ, and the maximum-likelihood filter is an estimator of θ.

Equation 1.269 can be viewed as either signal plus noise or $X(i)$ expressed as its mean θ plus its random displacement

$$N(i) = X(i) - \theta \tag{1.270}$$

from its mean. Under the latter interpretation, a set of observations is simply a random sample arising from an underlying random variable X, with each observation being identically distributed and maximum-likelihood estimation of θ being parametric estimation of the mean of X, as just discussed. Mathematically, nothing has changed; however, the interpretation has changed. Because our interest is in processing images and signals, we adopt the signal-model perspective for the remainder of this section.

In practice, a filter is applied to a signal or image by placing a *window* at a point and applying the filter to the observations in the window to estimate the signal value at the point. From this perspective, the signal of Eq. 1.269 is filtered by choosing a window containing some fixed number of points and applying the filter (estimator) to the values in the window at each location as the window is moved across the signal. Because the noise is the same across the entire signal in the model of Eq. 1.269, the same estimator is applied at each location. In this context, the sample mean is known as the *moving mean*.

***Example* 1.31.** The *Laplace distribution* is defined by the density

$$f(x) = \frac{\alpha}{2} e^{-\alpha|x-\mu|}$$

for all $x \in \Re$, and for parameters $\alpha > 0$ and $-\infty < \mu < \infty$. The Laplace distribution has mean μ and variance $2/\alpha^2$. Suppose $N(i)$ has a Laplace density with mean 0 and variance $2/\alpha^2$ in Eq. 1.269. Then the observations $X_1, X_2, \ldots, X_n$ form a random sample from a Laplace distribution with mean θ and variance $2/\alpha^2$. The likelihood function is

$$L(x_1, x_2, \ldots, x_n; \theta) = \left(\frac{\alpha}{2}\right)^n \exp\left[-\alpha \sum_{i=1}^{n} |x_i - \theta|\right]$$

Maximization occurs from minimizing the sum inside the exponential. Assuming the number of observations to be odd, this sum is minimized by letting $\hat{\theta}$ be the median of the observations; the resulting signal filter is called the *moving median*.

The maximum-likelihood estimator for additive Gaussian noise is the sample mean; the maximum-likelihood estimator for additive Laplacian noise is the sample median. Whatever the distribution, if $\mathrm{Var}[X] = \sigma^2$, then $\mathrm{Var}[\bar{X}] = \sigma^2/n$. For the sample median the matter is more complicated; however, for the Gaussian and Laplacian distributions, the sample median has asymptotic variances (as $n \to \infty$) $\pi\sigma^2/2n$ and $\sigma^2/2n$, respectively. Hence, for large n, the asymptotic variance of the sample median is approximately $\pi/2$ times as great as that for the sample mean when the underlying distribution is Gaussian, but the asymptotic variance of the sample median is approximately half as large as that for the sample mean when the underlying distribution is Laplacian. ■

***Example* 1.32.** Now assume in Eq. 1.269 that the noise is uniformly distributed over the interval $[-\beta, 0]$, where $\beta > 0$. The likelihood function is

$$L(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{\beta^n} \prod_{k=1}^{n} I_{[\theta-\beta,\theta]}(x_k)$$

where $I_{[\theta-\beta,\theta]}$ is the indicator function for the interval $[\theta - \beta, \theta]$. The likelihood function is

$$L(x_1, x_2, \ldots, x_n; \theta) = \begin{cases} \beta^{-n}, & \text{if } \theta - \beta \le x_k \le \theta \text{ for } k = 1,2,\ldots,n \\ 0, & \text{otherwise} \end{cases}$$

$L(\theta)$ is maximized for a set of sample values $x_1, x_2,..., x_n$, if and only if $\theta - \beta \leq x_k \leq \theta$ for all k, else $L(\theta) = 0$. Hence, $L(\theta)$ is maximized by having θ satisfy the inequality

$$\max\{x_1, x_2,..., x_n\} \leq \theta \leq \min\{x_1, x_2,..., x_n\} + \beta$$

If $\hat{\theta}$ is chosen so that

$$\max\{X_1, X_2,..., X_n\} \leq \hat{\theta} \leq \min\{X_1, X_2,..., X_n\} + \beta$$

then $\hat{\theta}$ is a maximum-likelihood estimator. If we assume β is unknown, then to be certain that the preceding double inequality holds, we must choose

$$\hat{\theta} = \max\{X_1, X_2,..., X_n\}$$

Applied across the entire signal, the maximum-likelihood filter is a *moving maximum*, which in morphological image processing is known as *flat dilation*.

If the model is changed with the noise uniformly distributed over the interval $[0, \beta]$, then the same basic analysis applies, except now $\hat{\theta}$ must be chosen so that

$$\max\{X_1, X_2,..., X_n\} - \beta \leq \hat{\theta} \leq \min\{X_1, X_2,..., X_n\}$$

If β is unknown, then we must choose

$$\hat{\theta} = \min\{X_1, X_2,..., X_n\}$$

which yields a *moving minimum* and is known as *flat erosion*. ■

***Example* 1.33.** Weighted medians are commonly employed in image processing. They give flexibility to the median approach by allowing certain observations to provide greater contributions to filter output. For the observations $x_1, x_2,..., x_n$ and the integer weights $\gamma_1, \gamma_2,..., \gamma_n$ (summing to an odd number), the *weighted median* for $x_1, x_2,..., x_n$ with weights $\gamma_1, \gamma_2,..., \gamma_n$ is defined to be the median with x_i repeated γ_i times for $i = 1, 2,..., n$. The weighted median can be viewed as a maximum-likelihood filter under the model of Eq. 1.269 if we drop the assumption that the noise observations are identically distributed and assume the noise for X_i is Laplacian distributed with mean 0 and variance $2/\gamma_i^2$. Then the likelihood function becomes

$$L(x_1, x_2,..., x_n; \theta) = \frac{\gamma_1\gamma_2\cdots\gamma_n}{2^n}\exp\left[-\sum_{i=1}^{n}\gamma_i|x_i-\theta|\right]$$

Maximization occurs with minimization of the sum inside the exponential. This sum can be expressed in the form of the sum in the exponent for the likelihood function for identically distributed Laplacian noise in Example 1.31. Just repeat each $|x_i - \theta|$ in that sum γ_i times. Hence the sum is minimized, and the likelihood function maximized, by the weighted median with weights $\gamma_1, \gamma_2,..., \gamma_n$. Since it has not been assumed that the noise is identically distributed, moving-window application is point dependent. ■

1.9.3. Minimum Noise

Rather than consider a constant signal corrupted by additive noise, as in Eq. 1.269, we can instead assume that the corrupted signal results from minimum noise; that is,

$$X(i) = \theta \wedge N(i) \tag{1.271}$$

where each $N(i)$ is identically distributed to a continuous random variable N. We assume that $F_N(\theta) < 1$; otherwise, we would have $N(i) \leq \theta$ with probability one. The probability distribution function for $X(i)$ is given by

$$F_{X(i)}(x) = P(X(i) \leq x) = \begin{cases} 1, & \text{if } x \geq \theta \\ F_N(x), & \text{if } x < \theta \end{cases} \tag{1.272}$$

Taking the derivative in a generalized sense yields $f_{X(i)}(x)$ and the likelihood function

$$L(x_1, x_2,..., x_n; \theta) = \prod_{k=1}^{n}[f_N(x_k)I_{(-\infty,\theta)}(x_k) + (1 - F_N(\theta))\delta(x_k - \theta)] \tag{1.273}$$

To find the maximum-likelihood filter, assume (without loss of generality) that $x_1, x_2,\ldots,x_q$ equal the maximum observation and $x_{q+1}, x_{q+2},\ldots,x_n$ are strictly less than the maximum. If

$$\theta = \max\{x_1, x_2,..., x_n\} \tag{1.274}$$

then

$$L(x_1, x_2,..., x_n; \theta) = \prod_{k=1}^{q} (1 - F_N(\theta))\delta(x_k - \theta) \prod_{k=q+1}^{n} f_N(x_k) \tag{1.275}$$

If $\theta < \max\{x_1, x_2,..., x_n\}$, then

$$f_N(x_1) I_{(-\infty,\theta)}(x_1) + (1 - F_N(\theta))\delta(x_1 - \theta) = 0 \tag{1.276}$$

so that $L(x_1, x_2,..., x_n; \theta) = 0$. If $\theta > \max\{x_1, x_2,..., x_n\}$, then

$$L(x_1, x_2,..., x_n; \theta) = \prod_{k=1}^{n} f_N(x_k) \tag{1.277}$$

Therefore, the maximum-likelihood filter is the maximum of the observations and, as a signal filter, is the flat dilation. A dual argument applies to maximum noise.

1.10. Entropy

Although we always lack certainty in predicting the outcome of a random variable, the degree of uncertainty is not the same in all cases. This section treats entropy, which is the quantification of uncertainty. Entropy plays a key role in coding theory, the subject of the next section.

1.10.1. Uncertainty

Consider a random variable X that can take on two values, 0 and 1, with probabilities $p = P(X = 1)$ and $q = P(X = 0)$. If $p = 0.99$ and $q = 0.01$, then the observer feels less uncertain than if the probabilities were $p = 0.6$ and $q = 0.4$. Intuitively, the observer's uncertainty is greatest when the probabilities are equal and least when one of the probabilities is 1. We now quantify uncertainty.

Suppose a discrete random variable X having probability mass function $f(x)$ can take on the n values $x_1, x_2,..., x_n$, with $f(x_i) = p_i$ for $i = 1, 2,..., n$. Because uncertainty should be increased when the probabilities p_i are more alike than when one or two carry most of the probability mass, the following criteria appear reasonable for a measure of uncertainty: (1) uncertainty is nonnegative and equal to zero if and only if there exists i such that $p_i = 1$; (2) uncertainty is maximum when the outcomes of X are equally likely; (3) if random variables X and Y have n and m equally likely outcomes, respectively, with $n < m$, then the uncertainty of X is less than the uncertainty of Y; and (4) uncertainty is a continuous function of $p_1, p_2,..., p_n$. These four conditions are met by defining the *uncertainty* of X by

$$H[X] = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1.278}$$

where the convention is adopted that $p_i \log_2 p_i = 0$ if $p_i = 0$. $H[X]$ is called the *entropy* of X; it is measured in *bits* and can be expressed in terms of expectation by

$$H[X] = -E[\log_2 f(X)] \tag{1.279}$$

The definition of entropy involves only probability masses, not their distribution on the x axis. Consequently, two random variables with the same outcome probabilities are indistinguishable from the perspective of entropy. Because total probability is 1, p_n can be considered to be a function of p_1, p_2,..., p_{n-1}, so that it is appropriate to write entropy as

$$H = H(p_1, p_2, \ldots, p_{n-1}) \tag{1.280}$$

We have claimed that $H[X]$ satisfies the four criteria listed for uncertainty. First, since $0 \le p_i \le 1$ for all i, it is immediately clear from the definition that $H[X] \ge 0$ and $H[X] = 0$ if and only if there exists i such that $p_i = 1$. Because $\log_2$ is a continuous function, H is also. If the outcomes of X are equally likely, then $H[X] = \log_2 n$ and entropy is an increasing function of n. To complete verification of the four criteria, we need to demonstrate that $H[X]$ is maximized in the case of equally likely outcomes. We treat the case of two outcomes having probabilities p and $1 - p$. Writing entropy as a function of p, assuming $0 < p < 1$, employing the natural logarithm, and differentiating yields

$$H'(p) = -\frac{1}{\log 2}\left[\log p - \log(1-p)\right] \tag{1.281}$$

Solving $H'(p) = 0$ yields $p = 0.5$. Since $H''(p) < 0$, a maximum occurs at $p = 0.5$.

Given two random variables X and Y, observing X can affect our uncertainty regarding Y. If X and Y can take on the values x_1, x_2,..., x_n and y_1, y_2,..., y_m, respectively, then, for $i = 1, 2, \ldots, n$, the *conditional entropy* of Y given x_i is defined via conditional densities by

$$H[Y|x_i] = -\sum_{j=1}^{m} f(y_j|x_i) \log_2 f(y_j|x_i) \tag{1.282}$$

This restates the entropy definition in terms of the conditional random variable $Y|x_i$. By letting x_i vary, conditional entropy becomes a function of X and is then written $H[Y|X]$.

As a function of X, $H[Y|X]$ has an expected value, called the *mean conditional entropy* of Y relative to X, and is defined by

$$\overline{H}[Y|X] = E[H[Y|X]]$$

$$= \sum_{i=1}^{n} H[Y|x_i] f_X(x_i) \tag{1.283}$$

Mean conditional entropy can be expressed via the conditional density of Y given X:

$$\overline{H}[Y|X] = -\sum_{i=1}^{n}\sum_{j=1}^{m} f_X(x_i) f(y_j|x_i) \log_2 f(y_j|x_i)$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, y_j) \log_2 f(y_j|x_i)$$

$$= -E[\log_2 f(Y|X)] \tag{1.284}$$

In particular, if X and Y are independent, then $\overline{H}[Y|X] = H[Y]$.

Generalizing Eq. 1.279 gives the definition of entropy for a random vector $(X, Y)'$:

$$H[X,Y] = -E[\log_2 f(X,Y)] \tag{1.285}$$

where $f(x, y)$ is the joint density of X and Y. A fundamental property is that the joint uncertainty of X and Y can be decomposed into a sum of the uncertainty of X plus the expected uncertainty remaining in Y following an observation of X, namely,

$$H[X,Y] = H[X] + \overline{H}[Y|X] \tag{1.286}$$

In particular, if X and Y are independent, then the joint uncertainty is the sum of the marginal uncertainties. Equation 1.286 is simply a restatement of the relation

$$-E[\log_2 f(Y|X) f_X(X)] = -E[\log_2 f(Y|X)] - E[\log_2 f_X(X)] \tag{1.287}$$

1.10.2. Information

Intuitively, if X and Y are dependent and X is observed, then information is obtained regarding Y because observation of X alters our uncertainty with respect to Y. The *expected amount of information* given for Y by observing X is defined by

$$I_X[Y] = H[Y] - \overline{H}[Y|X] \tag{1.288}$$

$I_X[Y]$ is the expected amount of information obtained, since, for a specific observation x_i of X, $H[Y] - H[Y|x_i]$ is the difference between the unconditional uncertainty of Y and the conditional uncertainty of Y given x_i, and $I_X[Y]$ results from taking the expected value of $H[Y] - H[Y|X]$ relative to X. If X and Y are independent, then $I_X[Y] = 0$.

Since the entropy of any random variable whose full probability mass is carried by a single outcome is 0 and since the conditional random variable $X|x_i$ is such a variable for any x_i, $H[X|x_i] = 0$. Hence, the mean conditional entropy of X relative to X is 0 and $I_X[X] = H[X]$, so that the entropy of X is the expected amount of information in X.

$I_X[Y]$ can be expressed via the joint and marginal densities of X and Y. From Eq. 1.288,

$$\begin{aligned} I_X[Y] &= -E[\log_2 f_Y(Y)] + E[\log_2 f(Y|X)] \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, y_j)\log_2 f(y_j) + \sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, y_j)\log_2 f(y_j|x_i) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, y_j)\left[\log_2 f(y_j|x_i) - \log_2 f(y_j)\right] \\ &= E\left[\log_2 \frac{f(Y|X)}{f_Y(Y)}\right] \\ &= E\left[\log_2 \frac{f(X,Y)}{f_X(X) f_Y(Y)}\right] \end{aligned} \tag{1.289}$$

Owing to the symmetry of the expression,

$$I_X[Y] = I_Y[X] \tag{1.290}$$

Using Eq. 1.289 and the fact that, for $a > 0$, $\log a \le a - 1$, we show that $I_X[Y] \ge 0$:

$$I_X[Y] = \sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, y_j) \log_2 \frac{f(x_i, y_j)}{f_X(x_i) f_Y(y_j)}$$

$$= -\frac{1}{\log 2}\sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, y_j) \log \frac{f_X(x_i) f_Y(y_j)}{f(x_i, y_j)}$$

$$\ge -\frac{1}{\log 2}\sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, y_j)\left[\frac{f_X(x_i) f_Y(y_j)}{f(x_i, y_j)} - 1\right]$$

$$= -\frac{1}{\log 2}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} f_X(x_i) f_Y(y_j) - \sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, y_j)\right] \tag{1.291}$$

Since both double sums in the preceding expression sum to 1, $I_X[Y] \ge 0$.

From the definition of $I_X[Y]$, it follows from Eq. 1.291 that

$$\overline{H}[Y|X] \le H[Y] \tag{1.292}$$

Applying this inequality to the decomposition of Eq. 1.286 yields

$$H[X, Y] \le H[X] + H[Y] \tag{1.293}$$

1.10.3. Entropy of a Random Vector

The entropy of an arbitrary random vector $(X_1, X_2, \ldots, X_r)'$ is defined by

$$H[X_1, X_2, \ldots, X_r] = -E[\log_2 f(X_1, X_2, \ldots, X_r)]$$

$$= -\sum_{x_1, x_2, \ldots, x_r} f(x_1, x_2, \ldots, x_r) \log_2 f(x_1, x_2, \ldots, x_r) \tag{1.294}$$

where $f(x_1, x_2, \ldots, x_r)$ is the joint density of $X_1, X_2, \ldots, X_r$. $H[X_1, X_2, \ldots, X_r]$ gives the expected information gained upon observation of $X_1, X_2, \ldots, X_r$. The basic properties holding for two random variables generalize to $r > 2$ random variables. In general,

$$H[X_1, X_2, \ldots, X_r] \le H[X_1] + H[X_2] + \cdots + H[X_r] \tag{1.295}$$

and, if $X_1, X_2,..., X_r$ are independent, then

$$H[X_1, X_2,..., X_r] = H[X_1] + H[X_2] + \cdots + H[X_r] \tag{1.296}$$

The mean conditional entropy for X_r given $X_1, X_2,..., X_{r-1}$ is defined by

$$\overline{H}\,[X_r \mid X_1, X_2,..., X_{r-1}] = -E[\log_2 f(X_r \mid X_1, X_2,..., X_{r-1})] \tag{1.297}$$

Corresponding to the decomposition of Eq. 1.286 is the decomposition

$$H[X_1, X_2,..., X_r] = H[X_1] + \sum_{k=2}^{r} \overline{H}[X_k \mid X_1, X_2, \ldots, X_{k-1}] \tag{1.298}$$

The inequality of Eq. 1.292 also extends: if $\{Z_1, Z_2,..., Z_s\} \subset \{X_1, X_2,..., X_{r-1}\}$, then

$$\overline{H}\,[X_r \mid X_1, X_2,..., X_{r-1}] \leq \overline{H}\,[X_r \mid Z_1, Z_2,..., Z_s] \tag{1.299}$$

For a sequence $\ldots, X_{-1}, X_0, X_1, \ldots$ of random variables, $H[X_{r-k}, X_{r-k+1},..., X_r]$ gives the uncertainty present in the $k + 1$ consecutive random variables terminating at time r. It follows from Eq. 1.299 that, for $k = 1, 2,...,$

$$\overline{H}\,[X_r \mid X_{r-1},..., X_{r-k}] \leq \overline{H}\,[X_r \mid X_{r-1},..., X_{r-k-1}] \tag{1.300}$$

Thus, $\overline{H}\,[X_r \mid X_{r-1},..., X_{r-k}]$ decreases as $k \to \infty$. Since $\overline{H}\,[X_r \mid X_{r-1},..., X_{r-k}]$ is bounded below by 0, it has a limit as $k \to \infty$. This limit,

$$\overline{H}_c[X_r] = \lim_{k\to\infty} \overline{H}[X_r \mid X_{r-1}, X_{r-2}, \ldots, X_{r-k}] \tag{1.301}$$

is called the (*mean*) *conditional entropy* of the sequence at time r and is a measure of the uncertainty concerning the present, X_r, given observation of the entire past.

1.11. Source Coding

For digital communication and storage, symbols have to be encoded into binary form. These symbols can be alphabetic, numeric, or a combination thereof; they can be image data, image features, or descriptive words or phrases pertaining to images. They are coded at a source and decoded at some destination. Source coding involves a finite set Σ of symbols, called the *source*, and the symbols (*source words*) need to be placed into one-to-one correspondence with strings (*code words*) of 0s and 1s. Efficient coding

requires that the expected number of bits in a randomly selected code word be kept small. To measure the efficiency of a particular encoding, we assume that the probability of each source word is known. If B is the random variable counting the number of bits transmitted for a randomly chosen source word, then the smaller the expected number of bits, $E[B]$, the more efficient the code. $E[B]$, which gives the expected length of a code word, does not depend on the actual symbols in Σ, but only on their codes and probabilities.

1.11.1. Prefix Codes

We consider only *prefix codes*. For these, no symbol code word can be obtained from the code word of a distinct symbol by adjoining 0s and 1s. For instance, if 001 is a code word, then 00110 cannot be the code word of a different symbol. A prefix code is uniquely decodable, meaning that a string of 0s and 1s forming a *message* (sequence of code words) can be unambiguously decoded by recognizing each code word in turn when reading from left to right without spacing between code words. For illustration, let $\Sigma = \{a, b, c, d\}$ and the code be given by the following correspondences: $a \leftrightarrow 01$, $b \leftrightarrow 001$, $c \leftrightarrow 10$, $d \leftrightarrow 110$. The message 01011011000100101110 is uniquely decoded as *aacdbbad*.

Since the actual source symbols are unimportant, we can assume without loss of generality that we are encoding the integers 1, 2,..., n. Thus, if p_1, p_2,..., p_n are the symbol probabilities, then in effect we have a random variable X taking on values 1, 2,..., n and having probability mass function $f(k) = p_k$ for $k = 1, 2,..., n$. If k is encoded using m_k bits, then the probability mass function for the code-word length B is defined, for $j = 1, 2,...$, by

$$f_B(j) = \sum_{\{k:m_k=j\}} p_k \tag{1.302}$$

Efficient coding requires using shorter code words for higher-probability symbols. Given n symbols, the problem is to design a prefix code for which $E[B]$ is minimal.

If each symbol probability is a negative power of 2, then the following coding algorithm yields a code possessing minimal expected code-word length:

C1. List the symbol probabilities in order from highest to lowest and label them from top to bottom as p_1, p_2,..., p_n. Ties can be broken arbitrarily.

C2. Beginning at the top of the list, select p_1, p_2,..., p_k so that

$$p_1 + p_2 + \cdots + p_k = 1/2$$

Let 1 be the first bit in the encoding of symbols 1 through k and 0 be the first bit in the encoding of symbols $k + 1$ through n.

C3. Out of the first k probabilities, choose the first r of them, p_1, p_2,..., p_r, so that

$$p_1 + p_2 + \cdots + p_r = 1/4$$

Let 1 be the second bit in the encoding of symbols 1 through r, so that their first two bits are 11, and let 0 be the second bit in the encoding of symbols $r + 1$ through k, so that their first two bits are 10. Similarly, choose the first s of the probabilities $k + 1$ through n, p_{k+1} through p_{k+s}, so that

$$p_{k+1} + p_{k+2} + \cdots + p_{k+s} = 1/4$$

Let 1 be the second bit in the encoding of symbols $k + 1$ through $k + s$, so that their first two bits are 01, and let 0 be the second bit in the encoding of symbols $k + s + 1$ through n, so that their first two bits are 00.

C4. Proceed recursively for the third and following bits, ceasing the process for any subcollection containing only a single symbol or for any subcollection containing two symbols after assigning 1 as the last bit of the encoding of the symbol higher on the list and 0 as the last bit of the encoding of the symbol lower on the list.

The coding algorithm generates a prefix code: if two bit strings are identical after generation of the rth bit, then they must represent symbols in the same subcollection at the rth stage of the algorithm, so that each string must get at least one more bit.

***Example* 1.34.** Consider the source Σ consisting of the first ten characters in the alphabet and the associated probabilities given in the following table:

Symbol	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
Probability	1/16	1/32	1/4	1/8	1/16	1/32	1/8	1/32	1/4	1/32

Application of the coding algorithm to Σ is illustrated in Table 1.1, together with the resulting code. According to the probabilities, $E[B] = 23/8$. Suppose we number the character rows of Table 1.1 and let X be the random variable that, for a selected character, gives its row number. Then a direct computation shows that $H[X] = E[B]$. ■

Table 1.1 Application of coding algorithm for Example 1.34.

Symbol	Prob	Bit Number					Code
		1	2	3	5	6	
c	1/4	1	1				11
i	1/4	1	0				10
d	1/8	0	1	1			011
g	1/8	0	1	0			010
a	1/16	0	0	1	1		0011
e	1/16	0	0	1	0		0010
b	1/32	0	0	0	1	1	00011
f	1/32	0	0	0	1	0	00010
h	1/32	0	0	0	0	1	00001
j	1/32	0	0	0	0	0	00000

The fact that $H[X] = E[B]$ in the preceding example is not exceptional; indeed, so long as the symbols are labeled by natural numbers in decreasing order of probability and all probabilities are negative powers of 2, the expected number of bits must equal the entropy of X when the coding algorithm is applied. To see this, suppose there are n symbols (n rows) and the probability corresponding to the kth row is $p_k = 2^{-m_k}$. According to the coding algorithm, the number of bits in the code of the symbol in the kth row is m_k. Thus,

$$E[B] = \sum_{k=1}^{n} m_k p_k$$

$$= \sum_{k=1}^{n} -p_k \log_2 p_k \tag{1.303}$$

Two questions arise. What happens when the symbol probabilities are not all negative powers of 2? And is it possible to find a more efficient coding scheme, one for which $E[B] < H[X]$? The coding algorithm provides a prefix code by assigning longer bit strings to symbols having lower probabilities. It is not optimal among all coding schemes, but it is optimal among prefix codes since, for these, $E[B] \geq H[X]$. When the probabilities are

all powers of 2, the coding algorithm provides an optimal prefix code in which the lower bound $H[X]$ is achieved. When the probabilities are not all powers of 2, the lower bound need not be achieved; however, there always exists a prefix code for which $E[B] - H[X] \leq 1$. Before demonstrating these statements, we examine possible code-word lengths in a prefix code, specifically, the possible encodings that result in r_j code words of length j.

For $j = 1$, a coding scheme can have 0, 1, or 2 code words of length 1. If $r_1 = 0$, then the final code has no encodings of the form $k \leftrightarrow 0$ or $k \leftrightarrow 1$; if $r_1 = 1$, then there exists a single encoding of the preceding types; and if $r_1 = 2$, then both $k \leftrightarrow 0$ and $l \leftrightarrow 1$ are in the code and, by the prefix requirement, they constitute the entire code. In all cases, $r_1 \leq 2$.

The number r_2 of code words of length $j = 2$ depends on r_1. If $r_1 = 2$, then the code is complete and $r_2 = 0$. If $r_1 = 1$, then r_2 can be 0, 1, or 2. To see this, suppose without loss of generality that the single code word of unit length is 1. Then $r_2 = 0$ if the code contains the code word 1 and other code words of the form $0b_1b_2\cdots b_m$ where $m \geq 2$; $r_2 = 1$ if the code contains the code words 1, 01, and other code words of the form $00b_1b_2\cdots b_m$ where $m \geq 1$; and $r_2 = 2$ if the full code consists of the code words 1, 01, and 00. Finally, if $r_1 = 0$, then r_2 can take on the values 0, 1, 2, 3, or 4. The case $r_2 = 4$ occurs when the full code is 11, 10, 01, and 00. The case $r_2 = 3$ occurs when the code words compose a set of the form 11, 10, 01, and further code words of the form $00b_1b_2\cdots b_m$, where $m \geq 1$. The case $r_2 = 2$ occurs when the code words compose a set of the form 11, 01, and further code words of the forms $00b_1b_2\cdots b_m$ and $10c_1c_2\cdots c_q$, where $m, q \geq 1$. The case $r_2 = 1$ occurs when the code words compose a set of the form 11 and further code words of the forms $00b_1b_2\cdots b_m$, $10c_1c_2\cdots c_q$, and $01d_1d_2\cdots d_s$ where $m, q, s \geq 1$. The case $r_2 = 0$ occurs when all code words are of the form $b_1b_2\cdots b_m$ with $m \geq 3$. To summarize, if $r_1 = 0$, then $r_2 \leq 4$; if $r_1 = 1$, then $r_2 \leq 2$; and if $r_1 = 2$, then $r_2 = 0$. These expressions may be combined to yield the single inequality $r_2 \leq 4 - 2r_1$.

Were we to continue, the number r_3 of code words of length $j = 3$ depends on r_1 and r_2. Indeed, it can be shown by mathematical induction that, for $j = 1, 2,\ldots, \nu$, where ν is the maximum of the code-word lengths,

$$r_j \leq 2^j - 2^{j-1}r_1 - 2^{j-2}r_2 - \cdots - 2r_{j-1} \tag{1.304}$$

Conversely, since the entire argument is constructive, if $r_1, r_2,\ldots, r_\nu$ is any sequence of nonnegative integers satisfying the inequality for $j = 1, 2,\ldots, \nu$, then there exists a prefix code having r_j code words of length $j = 1, 2,\ldots, \nu$.

***Example* 1.35.** To illustrate Eq. 1.304, let $r_1 = 1$, $r_2 = 0$, $r_3 = 3$, and $r_4 = 2$. Since $r_1 \leq 2$,

$$r_2 \leq 4 - 2r_1 = 2$$
$$r_3 \leq 8 - 4r_1 - 2r_2 = 4$$
$$r_4 \leq 16 - 8r_1 - 4r_2 - 2r_3 = 2$$

there exists a prefix code having 1, 0, 3, and 2 code words of lengths 1, 2, 3, and 4, respectively. One such code is composed of the code words 1, 011, 010, 001, 0001, and 0000. Another would be 1, 000, 011, 010, 0010, and 0011. ■

Theorem 1.22. For the source $\Sigma = \{1, 2,..., n\}$, there exists a prefix code for which the code word for k has m_k bits, $k = 1, 2,..., n$, if and only if

$$\sum_{k=1}^{n} 2^{-m_k} \leq 1 \qquad \blacksquare \qquad (1.305)$$

Equation 1.305 provides a lower bound on the sizes of the code-word lengths in a prefix code; however, since it is an equivalence, it has a valid converse. To demonstrate the meaning of the converse, suppose we are given code-word lengths $m_1 = 3$, $m_2 = 4$, $m_3 = 3$, $m_4 = 3$, and $m_5 = 2$. Since the sum in Eq. 1.305 yields 15/16, there exists a prefix code for the integers 1 through 5 with the given-word code lengths. One such code is given by $1 \leftrightarrow 101$, $2 \leftrightarrow 0001$, $3 \leftrightarrow 100$, $4 \leftrightarrow 011$, $5 \leftrightarrow 11$.

Theorem 1.22 is a consequence of Eq. 1.304, where in the context of the theorem, r_j gives the number of code-word lengths m_k equal to j. Owing to the relationship between $r_1, r_2,..., r_v$ and $m_1, m_2,..., m_n$,

$$\sum_{k=1}^{n} 2^{-m_k} = \sum_{j=1}^{v} r_j 2^{-j} \qquad (1.306)$$

Furthermore, collecting all the r terms on one side of Eq. 1.304 and dividing by 2^j yields the equivalent inequality

$$\sum_{i=1}^{j} r_i 2^{-i} \leq 1 \qquad (1.307)$$

Thus, the desired prefix code exists if and only if Eq. 1.307 holds for all j, which from Eq. 1.306 is true if and only if Eq. 1.305 holds.

Theorem 1.23. Let X be a random variable taking on the values 1, 2,..., n. If B counts the number of bits in a prefix encoding of $\{1, 2,..., n\}$, then

$$E[B] \geq H[X] \tag{1.308}$$

Moreover, there exists at least one prefix code such that

$$E[B] \leq H[X] + 1 \quad \blacksquare \tag{1.309}$$

To demonstrate that $H[X]$ is a lower bound for $E[B]$, let A denote the sum in Eq. 1.305, $p_k = P(X = k)$, and $t_k = 2^{-m_k}/A$. Then

$$\begin{aligned} -\sum_{k=1}^{n} p_k \log_2 \frac{p_k}{t_k} &= \log_2 e \sum_{k=1}^{n} p_k \log\left(\frac{t_k}{p_k}\right) \\ &\leq \log_2 e \sum_{k=1}^{n} p_k \left(\frac{t_k}{p_k} - 1\right) \\ &= \log_2 e \left(\sum_{k=1}^{n} t_k - \sum_{k=1}^{n} p_k\right) \end{aligned} \tag{1.310}$$

Since both sums in the last expression sum to 1, the sum on the left is bounded above by 0. Using this bound, expressing the logarithm of a quotient as the difference of the logarithms, and using the fact that $A \leq 1$, we demonstrate Eq. 1.308:

$$\begin{aligned} H[X] &= -\sum_{k=1}^{n} p_k \log_2 p_k \\ &\leq -\sum_{k=1}^{n} p_k \log_2 t_k \\ &= \sum_{k=1}^{n} p_k (m_k + \log_2 A) \\ &\leq \sum_{k=1}^{n} p_k m_k \end{aligned} \tag{1.311}$$

To fully prove Theorem 1.23, it remains to show that there exists at least one prefix code satisfying Eq. 1.309. For $k = 1, 2, \ldots, n$, let m_k be an integer for which

$$-\log_2 p_k \le m_k \le -\log_2 p_k + 1 \tag{1.312}$$

Then

$$\sum_{k=1}^{n} 2^{-m_k} \le \sum_{k=1}^{n} 2^{-\log_2 p_k} = 1 \tag{1.313}$$

and Theorem 1.22 ensures the existence of a prefix code with bit lengths m_k. Moreover,

$$\begin{aligned} E[B] &= \sum_{k=1}^{n} p_k m_k \\ &\le \sum_{k=1}^{n} p_k(-\log_2 p_k + 1) \\ &= H[X] + 1 \end{aligned} \tag{1.314}$$

***Example* 1.36.** This example elucidates the bit interpretation of entropy, mean conditional entropy, and expected amount of information in the context of prefix coding. Consider coding the sixteen numerals of the hexadecimal system assuming that their occurrences are equally likely. In effect, we have a random variable Y that can take on the values 0, 1, 2,..., 15 with equal probabilities. To obtain an optimal prefix code, we can apply the coding algorithm with the numerals listed ordinarily. Applying the coding algorithm and then interchanging the roles of 0 and 1 yields a code in which the code word of each hexadecimal integer is its binary equivalent: $0 \leftrightarrow 000$, $1 \leftrightarrow 0001$, $2 \leftrightarrow 0010$, $3 \leftrightarrow 0011$, $4 \leftrightarrow 0100$,..., $14 \leftrightarrow 1110$, $15 \leftrightarrow 1111$. Y has entropy $H[Y] = \log_2 16 = 4$. From an information perspective, the expected amount of information in Y is 4 bits, which is the number of bits required for the binary encoding of the hexadecimal system. Physically, transmission of a binary-encoded hexadecimal numeral requires 4 bits (so long as the probability mass is assumed to be equally distributed). Now suppose the numerals are grouped according to the first three bits of their binary codes, thereby creating eight classes numbered 0 through 7: $C_0 = \{0, 1\}$, $C_1 = \{2, 3\}$,..., $C_7 = \{14, 15\}$. If X gives the class number of an arbitrarily selected numeral, then X has eight equally likely outcomes and $H[X] = 3$. Given $X = i$, the equally likely outcomes of $Y|i$ mean that the conditional entropy of Y given $X = i$ is $H[Y|i] = \log_2 2 = 1$. Hence, the mean conditional entropy of Y relative to X is $\overline{H}[Y|X] = 1$ and the amount

of information obtained for Y by observing X is $I_X[Y] = 3$ bits, which agrees with X physically supplying 3 bits. In general, for $0 < s < r$, if Y has 2^r equally likely outcomes and X denotes the number of one of the 2^s classes resulting from equally dividing up the outcomes of Y, then $H[Y] = \log_2 2^r = r$, $\overline{H}[Y|X] = \log_2 2^{r-s} = r - s$, and $I_X[Y] = s$. The physical interpretation is that X has supplied s bits. The clarity of the interpretation results from the assumption that there are 2^r equally likely outcomes. In other circumstances, $I_X[Y]$ will not possess such a striking interpretation; nevertheless, the example illustrates the meaning of entropy and its relationship to information, as well as the underlying meanings of $H[Y]$, $\overline{H}[Y|X]$, and $I_X[Y]$. ■

1.11.2. Optimal Coding

Optimal coding involves finding a coding scheme for which the expected code-word length is minimized over all possible acceptable codes, in our case, binary prefix codes.

A prefix code can be viewed via a *coding tree*. Consider a tree, each branch being labeled 0 or 1, emanating from a root node and having a set of n terminal nodes. A prefix code for an n-symbol source is obtained by tracing out the branches from the root to each terminal node, associating a symbol with each terminal node, and assigning to each symbol the code word formed by the string of 0s and 1s running along the branches from the root to its corresponding terminal node. Moreover, every prefix code for a finite source can be so viewed.

A subclass of prefix codes is defined by trees for which each node is either terminal or has two branches (0 and 1) emanating from it. From the perspective of coding efficiency, nothing is lost by restricting ourselves to these *binary trees*. If d is a nonterminal node in a nonbinary tree from which there is generated only a single branch, then there must be a terminal node a subsequent to d in the tree so that the bit-strings to nodes d and a are of the forms $b_1 b_2 \cdots b_m$ and $b_1 b_2 \cdots b_m b_{m+1} \cdots b_q$, respectively. The code using

$$a \leftrightarrow b_1 b_2 \cdots b_m b_{m+1} \cdots b_q \tag{1.315}$$

can be replaced by a code that is exactly the same except that $a \leftrightarrow b_1 b_2 \cdots b_m$. The new code is a prefix code with $E[B]$ reduced. Subsequent to Theorem 1.22 we considered the code: $1 \leftrightarrow 101$, $2 \leftrightarrow 0001$, $3 \leftrightarrow 100$, $4 \leftrightarrow 011$, $5 \leftrightarrow 11$. A more efficient code results from "pruning" the coding tree back to obtain the code: $1 \leftrightarrow 101$, $2 \leftrightarrow 00$, $3 \leftrightarrow 100$, $4 \leftrightarrow 01$, $5 \leftrightarrow 11$, which is described by a binary coding tree. Since we are now concerned with efficiency, we henceforth assume that all codes correspond to binary coding trees.

When all symbol probabilities are negative powers of 2, an optimal code is achieved by the previously given coding algorithm with the lower bound

$H[X]$ being attained. More generally, Theorem 1.23 guarantees there exists a code whose expected bit length is within 1 of $H[X]$ but it does not provide a procedure for constructing an optimal code. The *Huffman code*, whose encoding algorithm we now describe, is optimal in the class of codes under consideration. To construct the Huffman code, list the source symbols according to their probabilities of occurrence, the highest probability down to the lowest. The coding tree is generated backward with the starting (terminal) nodes being the symbol-probability pairs $(a, p(a))$. Begin by combining two nodes of lowest probability, say $(a, p(a))$ and $(b, p(b))$, to form a new node $([a, b], p(a) + p(b))$. Branches of the coding tree run from node $([a, b], p(a) + p(b))$ to nodes $(a, p(a))$ and $(b, p(b))$. Label the top branch 0 and the bottom branch 1. The next node is generated by again combining two nodes of lowest probability, where the recently created node and its probability are part of the listing, and adding their probabilities. Branches connect the new node to the two chosen nodes and are labeled 0 and 1. Node creation is continued recursively until only a single node exists. The final node is the root and the original symbols are its terminal nodes. The Huffman code associates with each symbol the string of 0s and 1s going along the branches from the root node to the symbol. The choice of labeling the upper branch 0 and the lower branch 1 is arbitrary; at each stage, either can be labeled 0 and the other 1. Figure 1.3 shows a Huffman coding tree and the resulting code words.

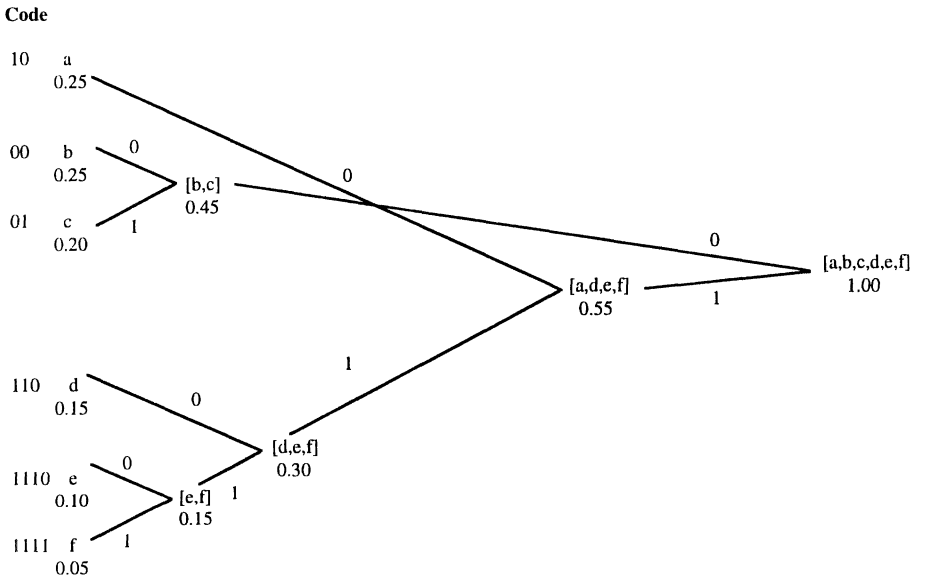

Figure 1.3 Huffman code.

Theorem 1.24. The Huffman code is optimal among prefix codes. ■

The theorem is demonstrated by mathematical induction on the number n of source words. If $n = 2$, then the Huffman code contains only two code words, 0 and 1, and is optimal. For the induction hypothesis, suppose the Huffman code is optimal for every source containing $m \leq n - 1$ symbols. To prove the theorem, we assume a source with n symbols, $\Sigma_n = \{a_1, a_2, \ldots, a_n\}$ with associated probabilities $p_1, p_2, \ldots, p_n$, and show that, based on the induction hypothesis, the Huffman code is optimal. Let β_n be the expected number of bits for the Huffman code and suppose there exists a different code, C_n, whose expected number of bits, α_n, is strictly less than β_n. For an optimal code, if the symbols a and b have probabilities $p(a)$ and $p(b)$ with $p(a) < p(b)$, then the length of the code word for a must be greater than or equal to the length of the code word for b. Consequently, if the symbols of Σ_n are listed in order of decreasing probability, it can be assumed without loss of generality that a_{n-1} and a_n appear as a node pair whose branches emanate from the same node at the last stage of the coding tree for C_n (for otherwise we could change the coding tree by interchanging symbols to produce a coding tree with all symbols having the same code-word lengths and with a_{n-1} and a_n appearing as a node pair whose branches emanate from the same node at the last stage of the tree). If the terminal nodes for symbols a_{n-1} and a_n are removed from the tree and a symbol c_{n-1} is placed at the new terminal node created, then the resulting tree provides a code C_{n-1} for the source $\Sigma_{n-1} = \{a_1, a_2, \ldots, a_{n-2}, c_{n-1}\}$. If the probability associated with c_{n-1} in Σ_{n-1} is $p_{n-1} + p_n$, then the expected number of bits for the new code C_{n-1} is

$$\alpha_{n-1} = \alpha_n - p_n - p_{n-1} \tag{1.316}$$

On the other hand, by construction of the Huffman code $\mathcal{H}_n$ for the n-symbol source Σ_n, a_{n-1} and a_n appear as a node pair whose branches emanate from the same node at the last stage of the coding tree for $\mathcal{H}_n$. If they are removed and replaced as in the case for C_n to form the source $\Sigma_{n-1} = \{a_1, a_2, \ldots, a_{n-2}, c_{n-1}\}$ with c_{n-1} having probability $p_{n-1} + p_n$, then the reduced tree is the coding tree for the Huffman code $\mathcal{H}_{n-1}$ corresponding to Σ_{n-1} and the expected number of bits for the Huffman code $\mathcal{H}_{n-1}$ is

$$\beta_{n-1} = \beta_n - p_n - p_{n-1} \tag{1.317}$$

Since, by supposition, $\alpha_n < \beta_n$, Eqs. 1.316 and 1.317 imply that $\alpha_{n-1} < \beta_{n-1}$; however, this contradicts the induction hypothesis and therefore we conclude that the Huffman code on n symbols is optimal.

Exercises for Chapter 1

1. Prove the probability addition theorem for $n = 3$ without using mathematical induction.
2. Show that, for $n \geq 1$ and $1 \leq k \leq n$, $C_{n,k} = C_{n-1,k-1} + C_{n-1,k}$.
3. Suppose there are m_k white balls and n_k black balls in urn A_k for $k = 1, 2,..., m$. Suppose an urn is randomly selected, with p_k being the probability of selecting urn A_k, and a ball is uniformly randomly selected from the chosen urn. Given that a black ball is selected, what is the probability that it was chosen from urn A_1?
4. Prove that, if E and F are independent, then so are E and F^c, E^c and F, and E^c and F^c.
5. Suppose an experiment has two possible outcomes, a and b, with the probability of a being $p > 0$ and the probability of b being $q > 0$. The experiment is independently repeated until the outcome a occurs and the random variable X counts the number of times b occurs before the outcome a. Find the probability density for X.
6. Find the constant c so the function $f(x)$ defined by $f(x) = cxe^{-bx}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$ is a legitimate density.
7. Demonstrate all the relations of Eq. 1.37.
8. Derive the density for $Y = |X|$ in terms of the density for X. Apply the result to the uniform density over $[-1, 2]$.
9. Derive the density for $Y = X^2$ in terms of the density for X.
10. Assuming $X \geq 0$, find the density for $Y = X^{1/2}$ in terms of the density for X. Apply the result to the case where X is uniformly distributed over $[4, 9]$.
11. Apply the result of Example 1.9 when X has an exponential density with parameter b.
12. The *Pareto distribution* has positive parameters r and a, and is defined by the density $f(x) = ra^r x^{-(r+1)}$ for $x \geq a$ and $f(x) = 0$ for $x < a$. Show that the Pareto distribution possesses a kth moment if and only if $k < r$. Assuming they exist, show that the mean and variance of the Pareto distribution are given by $\mu = ra/(r - 1)$ and $\sigma^2 = ra^2/(r - 1)^2(r - 2)$.
13. Derive the mean and variance of the exponential distribution directly from the density without using the moment-generating function.
14. Show that $\mathrm{Var}[X] = E[X^2] - E[X]^2$.
15. Show that $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$ for constants a and b.
16. Show that, for constants a and b, $M_{aX+b}(t) = e^{bt}M_X(at)$.
17. Compare the bound given by Chebyshev's inequality of Eq. 1.74 for the exponential distribution with the actual probability.
18. For the discrete random variable characterized by the density $f(-2) = f(2) = 1/8$ and $f(0) = 3/4$, show that Chebyshev's inequality is tight (cannot be improved).

19. If $f(x)$ is a continuous density, then its *median* is the value $\tilde{\mu}$ for which $P(X \le \tilde{\mu}) = P(X \ge \tilde{\mu})$. Find the median of the exponential and uniform densities.
20. For a continuous distribution, the *mean deviation* is defined by

$$MD(X) = \int_{-\infty}^{\infty} |x - \mu_X| f_X(x)\, dx$$

Find the mean deviation for the exponential and uniform distributions.
21. Find $M_X'(t)$ and $M_X''(t)$ for the binomial distribution. From these, derive μ, μ_2', and σ^2.
22. The *negative binomial distribution* is defined by the discrete density

$$f(x) = \binom{x+k-1}{k-1} p^k q^x$$

for $x = 0, 1, 2,...$, where k is a positive integer, $0 < p < 1$ and $p + q = 1$. Show that

$$M_X(t) = p^k(1 - qe^t)^{-k}$$

Using the moment-generating function, derive μ, μ_2', and σ^2.
23. For the Poisson density $\pi(x; \lambda)$ with parameter λ, show $\pi(0; \lambda) + \pi(1; \lambda) + \cdots = 1$.
24. Show that the Poisson density $\pi(x; \lambda)$ with parameter λ satisfies the recursion relation

$$\pi(x+1; \lambda) = \frac{\lambda}{x+1} \pi(x; \lambda)$$

25. Find $M_X'(t)$ and $M_X''(t)$ for the Poisson distribution. From these, derive μ, μ_2', and σ^2.
26. Using integration by parts, show that the probability distribution function of a Poisson random variable with parameter λ is given by

$$\Pi(x; \lambda) = \frac{1}{x!} \int_{\lambda}^{\infty} t^x e^{-t}\, dt$$

27. A discrete random variable X with density $f(x) = p(1 - p)^x$ for $x = 0, 1, 2,...$, where $0 < p < 1$, is called a *geometric* density (see Exercise 1.5). Show

$$M_X(t) = \frac{p}{1-(1-p)e^t}$$

Use the moment-generating function to show that $\mu = (1 - p)/p$ and $\sigma^2 = (1 - p)/p^2$.

28. Assuming the order of integration and summation can be interchanged, show that, for a continuous random variable X possessing finite moments of all orders,

$$M_X(t) = \sum_{k=0}^{\infty} E[X^k]\frac{t^k}{k!}$$

Assuming the order of summation can be interchanged, show the same result for a discrete random variable. Hint: In both instances, expand e^{tx} as a Maclaurin series.

29. Show that the total integral of the standard normal density is 1 (Eq. 1.98). Hint: Let I denote the integral, so that

$$I^2 = \left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\,dy\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dydx$$

and change to polar coordinates.

30. Find $M_X'(t)$, $M_X''(t)$, and $M_X'''(t)$ for the normal distribution. From these, derive μ, μ_2', μ_3', and σ^2.
31. In the text it was stated that the transformation $Z = (X - \mu)/\sigma$ transforms a normally distributed random variable X with mean μ and variance σ^2 into a standard normal variable and this is demonstrated by applying the result of Example 1.8. Show the details.
32. Use standard calculus maximum-minimum techniques to prove that the normal density has a maximum at μ and points of inflection at $\mu \pm \sigma$.
33. Find $M_X'''(t)$ for the gamma distribution and use it to find μ_3'.
34. For $\alpha = \nu/2$ and $\beta = 2$, the gamma distribution is known as the *chi-square distribution*. Write out the probability density, moment-generating function, mean, and variance for the chi-square distribution.
35. The *Weibull distribution* has density $f(x) = \alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^\alpha}$ for $x > 0$ and $f(x) = 0$ for $x \le 0$, where $\alpha > 0$ and $\beta > 0$. Show that the kth moment of the Weibull distribution is $\mu_k' = \beta^k\Gamma((\alpha + \beta)/\alpha)$. Hint: Use the substitution $u = (x/\beta)^\alpha$ in the integral giving $E[X^k]$. From μ_k' obtain the mean and variance.
36. Letting $\alpha = 2$ and $\beta = \sqrt{2}\,\eta$ in the Weibull distribution gives the *Rayleigh distribution* with parameter η (also see Example 1.19). From the

results of Exercise 1.35, find the mean and variance of the Rayleigh distribution.

37. Find the moment-generating function of the Laplace distribution and show that its mean and variance are given by μ and $2/\alpha^2$, respectively.
38. A basic operation in image processing is *thresholding*. For the input random variable X, the output $Y = g(X)$ is defined by $Y = 1$ if $X \geq \tau$ and $Y = 0$ if $X < \tau$, where τ is a fixed parameter. Show that, for X with a continuous density, the probability distribution function for Y is given by $F_Y(y) = 0$ for $y < 0$, $F_Y(y) = F_X(\tau)$ for $0 \leq y < 1$, and $F_Y(y) = 1$ for $y \geq 1$. Apply this result when X has a Laplace density with parameters μ and α.
39. Show that an exponentially distributed random variable is memoryless.
40. Consider the joint discrete densities

$$f(x,y) = \frac{3!}{x!\,y!(3-x-y)!\,2^x 3^y 6^{3-x-y}}$$

defined for all $x, y = 0, 1, 2, 3$ such that $x + y \leq 3$, and

$$g(x,y) = \binom{3}{x}\binom{3}{y}\frac{2^{3-y}}{216}$$

defined for all $x, y = 0, 1, 2, 3$. Numerically compute the outcome probabilities for both densities. Show that $g(x, y)$ is the product of two binomial densities, one with $n = 3$ and $p = 1/2$, and the other with $n = 3$ and $p = 1/3$. Show that both have the same marginal densities, these being the densities whose product gives $g(x, y)$.

41. Show parts (i) and (ii) of Theorem 1.10.
42. Repeat Example 1.16 with X and Y uniformly distributed over the region bounded by the curves $y = x^2$ and $y = x^{1/2}$, $x > 0$.
43. Suppose X and Y are uniformly distributed over the region bounded by the x axis, y axis, and the line $y = x/2 + 1$. Find $P(X < 1/2)$, $P(Y < X)$, and $P(Y > X^2)$.
44. Find the constant c so that the function $f(x, y) = ce^{-(2x + 4y)}$ for $x, y \geq 0$ and $f(x, y) = 0$ otherwise is a bivariate density. Find $P(X > Y)$ and the marginal densities.
45. Find the constant c so that the function $f(x, y) = cxe^{-(x + y)}$ for $x, y \geq 0$ and $f(x, y) = 0$ otherwise is a bivariate density. Find the marginal densities.
46. Let Z_1, Z_2, and Z_3 be independent standard normal random variables, $Y_1 = Z_1 + Z_2 + Z_3$, $Y_2 = Z_2 - Z_1$, and $Y_3 = Z_3 - Z_1$. Find the joint density of Y_1, Y_2, and Y_3.
47. Let X_1 and X_2 be independently distributed exponential random variables and find the density of $Y = X_1X_2$.

48. For $\kappa > 1$, find the covariance and correlation coefficient for the random variables X and Y possessing a uniform distribution over the region bounded by the curves $y = x^{\kappa}$ and $y = x^{1/\kappa}$, $x > 0$. Find the limits of the correlation coefficient as $\kappa \to 1$ and $\kappa \to \infty$.
49. Look up the Schwarz inequality and show how it relates to Theorem 1.14.
50. Show that if $Y = aX + b$, then $\mathrm{Cov}[X, Y] = a\mathrm{Var}[X]$.
51. Show that $\mathrm{Cov}[X, Y] = E[XY] - \mu_X\mu_Y$.
52. Show that the marginal distribution for X in the bivariate normal distribution is normally distributed with mean μ_X and variance σ_X^2 by integrating the joint normal density with respect to dy. Hint: let $u = (x - \mu_X)/\sigma_X$ and $v = (y - \mu_Y)/\sigma_Y$, complete the square in the quadratic in the resulting exponential, and make the change of variables $z = (1 - \rho^2)^{-1/2}(v - \rho u)$.
53. Suppose X possesses a Poisson distribution with parameter λ. Show that the limit as $\lambda \to \infty$ of the moment-generating function of the standardized Poisson random variable $(X - \lambda)/\lambda^{1/2}$ converges to the moment-generating function of the standard normal variable.
54. Rewrite Chebyshev's inequality in Eq. 1.200 for the case when $X_1, X_2, \ldots, X_n$ possess a multivariate normal distribution according to Eq. 1.190.
55. Suppose X is binomially distributed with parameters n and p. Based on the central limit theorem, write out an approximate expression for $P(a < X < b)$ in terms of the standard normal density under the supposition that n is large. Note that a rule-of-thumb is that the approximation is acceptable if $np \geq 5$ and $n(1 - p) \geq 5$.
56. Suppose X is Poisson distributed with mean λ. Based on the central limit theorem, write out an approximate expression for $P(a < X < b)$ in terms of the standard normal density under the supposition that n is large.
57. Apply the weak law of large numbers to the sequence of independent random variables $X_1, X_2, \ldots$ where X_n is Poisson distributed with mean λ when n is odd and binomially distributed with fixed parameters m and p when n is even.
58. Consider a sequence of uncorrelated random variables $X_1, X_2, \ldots$, where X_k has mean $k\alpha$, α fixed, and all variances are bounded by a common bound. Show that the weak law of large numbers states that, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{(n+1)\alpha}{2}\right| \geq \varepsilon\right) = 0$$

59. Show that the sample mean is an MVUE for the mean of the exponential distribution.

60. Show that the sample mean is an MVUE for the mean of the Poisson distribution.
61. Show that the sample mean is an MVUE for the mean of the binomial distribution with parameters $n = 1$ and p.
62. Find the method-of-moments estimator for the mean of the Poisson distribution.
63. Find the method-of-moments estimator for the mean of the geometric distribution.
64. Find the method-of-moments estimator for the parameter α of the beta distribution, given $\beta = 1$.
65. For X having an exponential distribution and $X_1, X_2, \ldots, X_n$ being a random sample arising from X, find the densities of the maximum and minimum order statistics.
66. Show that the median of a random sample of size $2n + 1$ from a uniform distribution over (0, 1) has a beta distribution with parameters $\alpha = \beta = n + 1$.
67. Suppose $X_1, X_2, \ldots, X_n, X_{n+1}, \ldots, X_m$ are independent identically distributed continuous random variables. Find

$$P(\min\{X_{n+1}, \ldots, X_m\} \geq \max\{X_1, X_2, \ldots, X_n\})$$

68. Using 1,000 random values generated by a uniform random-value generator, simulate a gamma distribution with parameters $\alpha = 20$ and $\beta = 0.5$. Construct the histogram of the data, use the sample mean and sample variance to obtain estimates of the mean and variance, and compare the estimates with the theoretical values.
69. Verify the last statement in Example 1.30, including the unbiasedness assertion.
70. Show that the sample mean is the maximum-likelihood estimator for the mean of the Poisson distribution.
71. Show that the sample mean is the maximum-likelihood estimator for the mean of the exponential distribution.
72. For a random sample of size n arising from a gamma distribution with α known and β unknown, show that $\hat{\beta} = \overline{X}/\alpha$ is the maximum-likelihood estimator for β.
73. Reconsider Example 1.32 with the noise uniformly distributed over $[0, \beta]$.
74. Reconsider Section 1.9.3 for maximum noise and show that the maximum-likelihood filter is a moving minimum (flat erosion).
75. Let X possess a uniform distribution over the interval $[\mu - 1/2, \mu + 1/2]$ and $Y_1, Y_2, \ldots, Y_n$ be the n order statistics corresponding to a random sample of odd size n. Find the expected values of the minimum Y_1 and the maximum Y_n. Suppose μ is unknown. Compare the sample-mean

estimator for μ with the estimator $\hat{\mu} = (Y_1 + Y_n)/2$. Specifically, check the bias and variance of $\hat{\mu}$.

76. In the text we showed that entropy is maximum for equally likely outcomes when there are two outcomes. Show the corresponding statement when there is an arbitrary number n of outcomes.
77. Demonstrate Eq. 1.295.
78. Demonstrate Eq. 1.298.
79. Demonstrate Eq. 1.299.
80. If the random variables ..., X_{-1}, X_0, X_1,... are independent and identically distributed, what can be said about the conditional entropy (Eq. 1.301)?
81. The entropy of a continuous random variable X possessing density $f(x)$ is defined by

$$H[X] = -E[\log_2 f(X)] = -\int_{-\infty}^{\infty} f(x) \log_2 f(x) dx$$

where the integration is assumed to be over $\{x: f(x) > 0\}$. Show that the entropy of a random variable uniformly distributed over the interval $(0, a)$ is $\log_2 a$.

82. Referring to the definition of Exercise 1.81, find the entropy of a normally distributed random variable possessing mean μ and variance σ^2.
83. Apply the coding algorithm of Section 1.11.1 to the source in the following table.

symbol	a	b	c	d	e	f	g	h	i
probability	1/8	1/32	1/8	1/8	1/8	1/32	1/8	1/16	1/4

Show that the lower entropy bound is achieved by the expected bit length.

84. The case for $j = 2$ in Eq. 1.304 is explained in detail in the text. Give a similar detailed explanation for $j = 3$.
85. Show that the code-word lengths $m_1 = 2$, $m_2 = 2$, $m_3 = 3$, $m_4 = 3$, and $m_5 = 3$ satisfy the condition of Theorem 1.22. Find all codes possessing these code-word lengths.
86. Find the Huffman code, $H[X]$, and $E[B]$ for the source in the following table:

symbol	a	b	c	d	e	f	g	h	i
probability	1/10	1/10	1/20	1/5	1/5	1/20	1/20	3/20	1/10

87. Constructing the Huffman code can be complicated when there is a large number of symbols. To avoid cumbersome coding, with the loss of some efficiency, one can employ a *truncated Huffman code*. For truncation, the symbols with the smallest probabilities are grouped into a single group symbol for the purpose of the Huffman code tree. When the tree is completed and coding accomplished, symbols forming the group symbol are individually coded by appending a fixed-bit-length binary code to the code word generated by the coding tree. If there are between 5 and 8 grouped symbols, then 3 bits are appended; if there are between 9 and 16 grouped symbols, then 4 bits are appended; etc. Apply truncated Huffman coding to the following source by combining the symbols with the six smallest probabilities into a group:

symbol	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*
probability	0.025	0.015	0.050	0.050	0.075	0.010	0.020	0.025

symbol	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*
probability	0.075	0.125	0.075	0.175	0.005	0.050	0.100	0.125

88. Truncate the Huffman code for the source of Exercise 1.86 by grouping the symbols with the three lowest probabilities. Compare the expected bit lengths for the Huffman and truncated Huffman codes for this source.
89. Suppose a_1 and a_2 are two source symbols with probabilities p_1 and p_2, respectively, the source is coded optimally, and the code words for a_1 and a_2 are of lengths m_1 and m_2, respectively. Show that, if $p_1 < p_2$, then $m_1 \geq m_2$.