APPENDIX A: PROBABILITY

Contents

1

Basic Probability Models

Further details concerning the first section of the appendix can be found in most introductory texts in probability and mathematical statistics. The material in the second and third chapters can be supplemented with Steele (2001) for further details and many of the proofs.

1.1 BASIC DEFINITIONS

Probabilities are defined on sets or events, usually denoted with capital letters early in the alphabet such as A, B, C. These sets are subset of a *sample space* or *probability space* Ω, which one can think of as a space or set containing all possible outcomes of an experiment. We will say that an event $A \subset \Omega$ occurs if one of the outcomes in A (rather than one of the outcomes in Ω but outside of A) occurs. Not only should we be able to describe the probabilities of individual events, we should also be able to define probabilities of various combinations of them, including

1. Union of sets or events: $A \cup B = A$ or B (occurs whenever A occurs or B occurs or both A and B occur)
2. Intersection of sets: $A \cap B = A$ and B (occurs whenever A and B occur)
3. Complement: A^c = not A (occurs when the outcome is not in A)
4. Set differences: $A \setminus B = A \cap B^c$ (occurs when A occurs but B does not)
5. Empty set: $\phi = \Omega^c$ (an impossible event—it never occurs since it contains no outcomes)

Recall *De Morgan's rules* of set theory: $(\cup_i A_i)^c = \cap_i A_i^c$ and $(\cap_i A_i)^c = \cup_i A_i^c$.

Events are subsets of Ω. We will call $\mathcal{F}$ the class of all events (including ϕ and Ω).

Definition A *probability measure* is a set function $P : \mathcal{F} \to [0, 1]$ such that

6. $P(\Omega) = 1$

7. If A_k is a disjoint sequence of events so that $A_k \cap A_j = \phi$, for $k \neq j$, then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

These are the basic axioms of a probability model. From these it is not difficult to prove the following properties:

1. $P(\phi) = 0$.
2. If $A_k, k = 1, \ldots, N$, is a finite or countable sequence of disjoint events so that $A_k \cap A_j = \phi, k \neq j$, then

$$P(\cup_{i=1}^{N} A_i) = \sum_{i=1}^{N} P(A_i)$$

3. $P(A^c) = 1 - P(A)$.
4. Suppose $A \subset B$. Then $P(A) \leq P(B)$.
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
6. The inclusion-exclusion principle:

$$P(\cup_k A_k) = \sum_k P(A_k) - \sum\sum_{i<j} P(A_i \cap A_j) + \sum\sum\sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots .$$

7. $P(\cup_{i=1}^{\infty} A_i) \leq \sum_i P(A_i)$.
8. Suppose $A_1 \subset A_2 \subset \cdots$. Then $P(\cup_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i)$.

Counting Techniques

Permutations The number of ways of permuting or arranging n distinct objects in a row is $n! = n(n-1)\cdots 1$ and $0! = 1$. Define $n^{(r)} = n(n-1)\cdots(n-r+1)$ (called "n to r factors") for arbitrary n, and r a nonnegative integer. This is the number of permutations of n objects taken r at a time. Define $n^{(0)} = 1$ and notice that values like $(\frac{1}{2})^{(3)}$ are well defined (indeed, $(\frac{1}{2})^{(3)} = (\frac{1}{2})(-\frac{1}{2})(-\frac{3}{2}) = \frac{3}{8}$).

For example, the number of distinct ways of rearranging the 15 letters

$$AAAAABBBBCCCDDE$$

would be 15! if all 15 letters could be distinguished. Since they cannot, this calculation counts the two possible orderings of the Ds (e.g., D_1D_2 or D_2D_2)

separately, and the 3! reorderings of the *Cs* are counted separately, etc. Therefore, dividing by the number of times each letter has been overcounted, the number of distinct rearrangements is

$$\frac{15!}{5!4!3!2!} = \binom{15}{5\quad 4\quad 3\quad 2}$$

Combinations Suppose the order of selection is not considered to be important. We wish, for example, to distinguish only different *sets* selected, without regard to the order in which they were selected. Then the number of distinct sets of r objects that can be constructed from n distinct objects is

$$\binom{n}{r} = \frac{n^{(r)}}{r!}$$

Note this is well defined for r a nonnegative integer for any real number n.

Independent Events Two events A, B are said to be *independent* if

$$P(A \cap B) = P(A)P(B) \tag{1.1}$$

Compare this definition with that of *mutually exclusive* or *disjoint* events A, B. Events A, B are mutually exclusive if $A \cap B = \phi$.

Independent experiments are often built from *Cartesian products* of sample spaces. For example, if Ω_1 and Ω_2 are two sample spaces, and $A_1 \subset \Omega_1, A_2 \subset \Omega_2$, then an experiment consisting of *both of the above* would have as sample space the Cartesian product

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2); \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$$

Probabilities of events such as $A_1 \times A_2$ are easily defined, in this case as $P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$. Verify in this case that an event entirely determined by the first experiment such as $A = A_1 \times \Omega_2$, is independent of one determined by the second, $B = \Omega_1 \times A_2$.

Definition A finite or countably infinite set of events $A_1, A_2, \ldots$ are said to be mutually independent if

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}) \tag{1.2}$$

for any $k \geq 2$ and $i_1 < i_2 < \cdots < i_k$.

Independent events have the properties that

1. A, B independent implies A, B^c independent.

2. Any A_{i_j} can be replaced by $A_{i_j}^c$ in equation (1.2).

Why not simply require that every pair of events be independent? This is, as it turns out, too weak an assumption for many of the results we need in probability and statistics, and does not describe what we intuitively mean by independence. For example, suppose two fair coins are tossed. Let $A =$ first coin is heads, $B=$ second coin is heads, $C=$ we obtain exactly one heads. Then A is independent of B and A is independent of C, but A, B, C are **not mutually independent**. Thus *pairwise independence does not imply independence.* Does it make intuitive sense to say that A, B, C are independent? If you know whether A and B occur, then you automatically know whether or not the event C occurs, so there is a strong dependence among these three events.

"Lim Sup" of events For a sequence of events $A_n, n = 1, 2, \ldots$, we define another event $[A_n\ i.o.] = \limsup A_n = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n$. Note that this is the set of all points x that lie in infinitely many of the events $A_1, A_2, \ldots$ The notation *i.o.* stands for "infinitely often" because $\limsup A_n$ is the set of all points ω that are in infinitely many of the $A_n, n = 1, 2, \ldots$. There is a similar notion, $\liminf A_n = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n$, and it is not difficult to show that the latter set is smaller:

$$\liminf A_n \subset \limsup A_n$$

A point ω is in $\liminf A_n$ if and only if it is in all of the sets A_n except possibly a finite number. For this reason we sometimes denote $\liminf A_n$ as $[A_n\ a.b.f.o.]$, where *a.b.f.o.* stands for "all but finitely often."

Borel Cantelli Lemmas Clearly, if events are individually too small, there is little or no probability that their lim sup will occur (i.e., that they will occur infinitely often). This is the essential message of the first of the Borel-Cantelli lemmas:

Lemma A1 *For an arbitrary sequence of events* A_n, *if* $\sum_n P(A_n) < \infty$ *then* $P[A_n\ i.o.] = 0$.

Lemma A2 *For a sequence of* independent events A_n, $\sum_n P(A_n) = \infty$ *implies* $P[A_n\ i.o] = 1$.

Conditional Probability Suppose we are interested in the probability of the event A but we are given some relevant information, namely that another, related event B occurred. How do we revise the probabilities assigned to points of Ω in view of this information? If the information does not affect the relative probability of points in B, then the new probabilities of points outside of B should be set to 0 and those within B simply rescaled to add to one. This is essentially the definition of conditional probability given B. Given that B occurred, reassign probability 0 to those points outside of B and rescale those within so that the total probability is 1.

Definition: Conditional Probability For $B \in \mathcal{F}$ with $P(B) > 0$, define a new probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.3}$$

This is also a probability measure on the same space $(\Omega, \mathcal{F})$ and satisfies the same properties. Note that $P(B|B) = 1, P(B^c|B) = 0$.

Theorem A1 (Bayes' Rule) *If $P(\cup_n B_n) = 1$ for a disjoint finite or countable sequence of events B_n all with positive probability, then*

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_n P(A|B_n)P(B_n)} \tag{1.4}$$

Theorem A2 (Multiplication Rule) *If $A_1, \ldots, A_n$ are arbitrary events,*

$$P(A_1A_2 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2A_1) \cdots P(A_n|A_1A_2 \cdots A_{n-1}) \tag{1.5}$$

Random Variables

Properties of $\mathcal{F}$ The class of events $\mathcal{F}$ (called a sigma algebra or sigma field) should be such that the operations normally conducted on events, for example, countable unions or intersections, or complements, keeps us within that class. In particular,

(a) $\varphi \in \mathcal{F}$
(b) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$
(c) If $A_n \in \mathcal{F}$ for all $n = 1, 2, \ldots$), then $\cup_{n=1}^{\infty} \in \mathcal{F}$

It follows from these properties that $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is also closed under countable intersections, countable intersections of unions, and so on.

Definition Let X be a function from a probability space Ω into the real numbers. We say that the function is *measurable* (in which case we call it a random variable) if for $x \in \Re$, the set $\{\omega; X(\omega) \leq x\} \in \mathcal{F}$. Since events in $\mathcal{F}$ are those to which we can attach a probability, this permits us to obtain probabilities for the event that the random variable X is less than or equal to any number x.

Definition: Indicator Random Variables For an arbitrary set $A \in \mathcal{F}$ define $I_A(\omega) = 1$ if $\omega \in A$ and 0 otherwise. This is called an *indicator random variable* (sometimes called a *characteristic function* in measure theory, but not here).

Definition: Simple Random Variables Consider events $A_i \in \mathcal{F}$ such that $\cup_i A_i = \Omega$. Define $X(\omega) = \sum_{i=1}^{n} c_i I_{A_i}(\omega)$, where $c_i \in \Re$. Then X is measurable and is consequently a random variable. We normally assume that the sets A_i are disjoint. Because this is a random variable that can take only finitely many different values, it is called *simple*. Any random variable taking only finitely many possible values can be written in this form.

We will often denote the event $\{\omega \in \Omega; X(\omega) \leq x\}$ more compactly by $[X \leq x]$. In general, functions of one or more random variables gives us another random variable (provided that function is measurable). For example, if X_1, X_2 are random variables, so are

1. $X_1 + X_2$
2. $X_1 X_2$
3. $\min(X_1, X_2)$.

The *cumulative distribution function* of a *random variable* X is defined to be the function $F(x) = P[X \leq x]$, for $x \in \Re$.

Properties of the Cumulative Distribution Function

1. A cumulative distribution function $F(x)$ is nondecreasing (i.e., $F(x) \geq F(y)$ whenever $x \geq y$).
2. $F(x) \to 0$, as $x \to -\infty$.
3. $F(x) \to 1, x \to \infty$.
4. $F(x)$ is right continuous: $F(x) = \lim_{h \to 0^+} F(x+h)$ (i.e., the limit as h decreases to 0).

There are two primary types of distributions considered here, discrete distributions and continuous ones. Discrete distributions are those whose cumulative distribution function at any point x can be expressed as a finite or countable sum of values. For example,

$$F(x) = \sum_{\{i; x_i \leq x\}} p_i$$

for some probabilities p_i that sum to 1. In this case the cumulative distribution is piecewise constant, with jumps at the values x_i that the random variable can assume. The values of those jumps are the individual probabilities p_i. For example $P[X = x]$ is equal to the size of the jump in the graph of the cumulative distribution function at the point x. We refer to the function $f(x) = P[X = x]$ as the *probability function* of the distribution when the distribution is discrete.

1.2 SOME SPECIAL DISCRETE DISTRIBUTIONS

The Discrete Uniform Distribution Many of the distributions considered so far are such that each point is equally likely. For example, suppose the random variable X takes each of the points $a, a+1, \ldots, b$ with the same probability $\frac{1}{b-a+1}$. Then the cumulative distribution function is

$$F(x) = \frac{x-a+1}{b-a+1}, \quad x = a, a+1, \ldots, b$$

and the probability function is $f(x) = \frac{1}{b-a+1}$ for $x = a, a+1, \ldots, b$ and 0 otherwise.

The Hypergeometric Distribution Suppose we have a collection (the *population*) of N objects that can be classified into two groups S and F where there are r of type S and $N-r$ of type F. Suppose we take a random sample of n items without replacement from this population. Then the probability that we obtain exactly x items of type S is

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, \ldots$$

The Binomial Distribution The setup is identical to that in the last paragraph, only now we sample *with replacement*. Thus, for each member of the sample, the probability of an S is $p = r/N$. Then the probability function is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots n$$

With any distribution, the sum of *all* the probabilities should be 1. Check that this is the case for the binomial,

$$\sum_{x=0}^{n} f(x) = 1$$

The hypergeometric distribution is often approximated by the binomial distribution in the case N large. For the binomial distribution, the two *parameters* n, p are fixed, and usually known. For fixed sample size n we count $X =$ "number of S's in n *trials* of a simple experiment" (e.g., tossing a coin).

The Negative Binomial Distribution The binomial distribution was generated by assuming that we repeated trials a fixed number n of times and then counted the total number of successes X in those n trials. Suppose we decide in advance that we wish a fixed number (k) of successes instead, and sample repeatedly until we obtain exactly this number. Then the number of trials X is random.

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \ldots$$

A special case of interest is the case $k = 1$, called the *geometric* distribution. Then

$$f(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

The Poisson Distribution. Suppose a disease strikes members of a large population (of n individuals) independently, but in each case it strikes with very small probability p. If we count X, the number of cases of the disease in the population, then X has the binomial(n, p) distribution. For very large n and small p this distribution can be again approximated as follows:

Theorem A3 *Suppose $f_n(x)$ is the probability function of a binomial distribution with $p = \lambda/n$ for some fixed λ. Then as $n \to \infty$,*

$$f_n(x) \to f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for each $x = 0, 1, 2, \ldots$.

The function $f(x)$ above is the probability function of a *Poisson distribution*, named after a French mathematician. This distribution has a single parameter, λ, which makes it easier to use than the binomial, since the binomial requires knowledge or estimation of two parameters. For example, the size n of the population of individuals who are susceptible to the disease might be unknown, but the "average" number of cases in a population of a certain size might be obtainable from medical data.

1.3 EXPECTED VALUE

An indicator random variable I_A takes two values, the value 1 with probability $P(A)$ and the value 0 otherwise. Its expected value, or average over many (independent) trials, would therefore be $0(1-P(A))+1P(A)=P(A)$. This is the simplest case of an integral or expectation.

Recall that a simple random variable is one that has only finitely many distinct values c_i on the sets A_i, where these sets form a partition of the sample space (i.e., they are disjoint and their union is Ω).

Expectation of Simple Random Variables For a simple random variable $X = \sum_i c_i I_{A_i}$, define $E(X) = \sum_i c_i P(A_i)$. The form is standard:

$$E(X) = \sum (\text{values of } X) \times (\text{probability of values})$$

Thus, for example, if a random variable X has probability function $f(x) = P[X = x]$, then $E(X) = \sum_x xf(x)$.

Example The expected value of X, a random variable having the binomial(n, p) distribution, is $E(X) = np$.

Expectation of Nonnegative Measurable Random Variables

Definition Suppose X is a nonnegative random variable so that $X(\omega) \geq 0$ for all $\omega \in \Omega$. Then we define

$$E(X) = \sup\{E(Y);\ Y \text{ simple and } Y \leq X\}$$

Expected Value: Discrete Case If a random variable X has probability function $f(x) = P[X = x]$, then the definition of expected value in the case of *finitely many* possible values of x is essentially $E(X) = \sum_x xf(x)$. This formula continues to hold even when X may take a countably infinite number of values provided that the series $\sum_x xf(x)$ is absolutely convergent.

Notation Note that by $\int_A X\,dP$ we mean $E(XI_A)$, where I_A is the indicator of the event A.

Properties of Expectation Assume X, Y are nonnegative random variables. Then

1. If $X = \sum_i c_i I_{A_i}$ is simple, then $E(X) = \sum_i c_i P(A_i)$.
2. If $X(\omega) \leq Y(\omega)$ for all ω, then $E(X) \leq E(Y)$.

3. If X_n is increasing to X, then $E(X_n)$ increases to $E(X)$ (this is usually called the *monotone convergence theorem*).
4. For nonnegative numbers α, β, $E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.

General Definition of Expected Value For an arbitrary random variable X, define two new random variables $X^+ = \max(X, 0)$, and $X^- = \max(0, -X)$. Note that $X = X^+ - X^-$. Then we define $E(X) = E(X^+) - E(X^-)$. This is well defined even if one of $E(X^+)$ or $E(X^-)$ is equal to ∞ as long as both are not infinite since the form $\infty - \infty$ is meaningless. If both $E(X^+) < \infty$ and $E(X^-) < \infty$, then we say X is *integrable.*

Example Define a random variable X such that $P[X = x] = \frac{1}{x(x+1)}$, $x = 1, 2, \ldots$ Is this random variable integrable? If we write out the expected value,

$$\sum_{x=1}^{\infty} xf(x) = \sum_{x=1}^{\infty} \frac{1}{x+1}$$

and this is a divergent sequence, so in this case the random variable is not integrable.

General Properties of Expectation In the general case, expectation satisfies 1–4 above plus the additional properties

5. If $P(A) = 0$, then $\int_A X(\omega) dP = 0$.
6. If $P[X = c] = 1$ for some constant c, then $E(X) = c$.
7. If $P[X \geq 0] = 1$ then $E(X) \geq 0$.

Other Interpretations of Expected Value For a discrete distribution, the distribution is often represented graphically with a bar graph or histogram. If the values of the random variable are $x_1 < x_2 < x_3 < \cdots$, then rectangles are constructed around each value, x_i, with *area* equal to the probability $P[X = x_i]$. In the usual case where the x_i are equally spaced, the rectangle around x_i has as base $(\frac{x_{i-1}+x_i}{2}, \frac{x_i+x_{i+1}}{2})$. In this case, the expected value $E(X)$ is the x-coordinate of the center of gravity of the probability histogram.

We may also think of expected value as a long-run average over many independent repetitions of the experiment. Thus, $f(x) = P[X = x]$ is approximately the long-run proportion of occasions on which we observed the value $X = x$, so the *long-run average* of many independent replications of X is $\sum_x xf(x) = E(X)$.

Lemma (Fatou's Lemma: Limits of Integrals) *If X_n is a sequence of nonnegative random variables,*

$$E[\liminf X_n] \leq \liminf EX_n$$

It is possible that $X_n(\omega) \to X(\omega)$ for all ω and yet $E(X_n)$ does not converge to $E(X)$. For example, let $\Omega = (0, 1)$ and the probability measure be the Lebesgue measure on the interval. Define $X(\omega) = n$ if $0 < \omega < 1/n$ and otherwise $X(\omega) = 0$. Then $X_n(\omega) \to 0$ for all ω, but $E(X_n) = 1$ does not converge to the expected value of the limit. This example shows that some additional condition is required beyond (almost sure) convergence of the random variables in order to conclude that the expected values converge. One such condition is given in the following important result.

Theorem A4 (Lebesgue-Dominated Convergence Theorem) *If $X_n(\omega) \to X(\omega)$ for each ω, and there exists integrable Y with $|X_n(\omega)| \leq Y(\omega)$ for all n, ω, then X is integrable and $E(X_n) \to E(X)$.*

Lebesgue-Stieltjes Integral

A basic requirement of any sigma algebra of subsets of the real line for it to be of much use is that it contain the intervals, since we often wish to compute probabilities of intervals like $[a < X < b]$.

Definition: Borel Sigma Algebra The smallest sigma algebra that contains all of the open intervals is called the Borel sigma algebra. The sets in this sigma algebra are referred to as Borel sets.

Fortunately it is easy to show that this sigma algebra also contains all of the closed intervals—in fact, all countable unions of intervals of any kind, open, closed, or half open. We call a function $g(x)$ on the real numbers (i.e., $\Re \to \Re$) Borel measurable if for any Borel subset $B \subset \Re$, the set $\{x; g(x) \in B\}$ is also a Borel set.

We now consider integration of functions on the real line or Euclidean space. Suppose $g(x)$ is a Borel measurable function $\Re \to \Re$. Suppose $F(x)$ is a Borel measurable function satisfying

1. $F(x)$ is nondecreasing (i.e., $F(x) \geq F(y)$ whenever $x \geq y$).
2. $F(x)$ is right continuous (i.e., $F(x) = \lim F(x + h)$ as h decreases to 0).

Notice that we can use F to define a measure μ on the real line; for example, the measure of the interval $(a, b]$ we can take to be $\mu((a, b]) = F(b) - F(a)$. The measure is extended from these intervals to all Borel sets in the usual way, by first defining the measure on the algebra of finite unions of intervals, and then extending this measure to the Borel sigma algebra generated by this algebra. We will define $\int g(x)dF(x)$ or $\int g(x)d\mu$ exactly as we did expected values, but with the probability measure P replaced by μ and $X(\omega)$ replaced by $g(x)$. In particular, for a simple function $g(x) = \sum_i c_i I_{A_i}(x)$, we define $\int g(x)dF = \sum_i c_i \mu(A_i)$.

Definition: Integration of Borel Measurable Functions Suppose $g(x)$ is a nonnegative Borel measurable function so that $g(x) \geq 0$ for all $x \in \Re$. Then we define

$$\int g(x)d\mu = \sup\left\{\int h(x)d\mu;\, h \text{ is simple and } h \leq g\right\}$$

For a general function $f(x)$ we write $f(x) = f^+(x) - f^-(x)$ where both f^+ and f^- are nonnegative functions. We then define $\int f\, d\mu = \int f^+\, d\mu - \int f^-\, d\mu$ provided that this makes sense (i.e., is not of the form $\infty - \infty$). Finally, we say that f is integrable if both f^+ and f^- have *finite integrals*, or equivalently, if $\int |f(x)|d\mu < \infty$.

Definition: Absolutely Continuous A measure μ on $\Re$ is *absolutely continuous* with respect to Lebesgue measure λ (denoted $\mu \ll \lambda$) if there is an integrable function $f(x)$ such that $\mu(B) = \int_B f(x)d\lambda$ for all Borel sets B. The function f is called the *density* of the measure μ with respect to λ.

Intuitively, two measures μ, λ on the same measurable space $(\Omega, \mathcal{F})$ (not necessarily the real line) satisfy $\mu \ll \lambda$ if the support of the measure λ includes the support of the measure μ. For a discrete space, the measure μ simply reweights those points with nonzero probabilities under λ. For example, if λ represents the discrete uniform distribution on the set $\Omega = \{1, 2, 3, \ldots, N\}$ (so that $\lambda(B)$ is N^{-1} × the number of integers in $B \cap \{1, 2, 3, \ldots, N\}$) and $f(x) = x$, then if $\mu(B) = \int_B f(x)d\lambda$, we have $\mu(B) = \sum_{x \in B \cap \{1,2,3,\ldots,N\}} x$. Note that the measure μ assigns weights $\frac{1}{N}, \frac{2}{N}, \ldots, 1$ to the points $\{1, 2, 3, \ldots, N\}$, respectively.

The so-called continuous distributions, such as the normal, gamma, exponential, beta, chi-squared and Student's t, should be called *absolutely continuous with respect to Lebesgue measure* rather than just continuous.

Theorem A5 (Radon-Nykodym Theorem) *For arbitrary measures μ and λ defined on the same measure space, the two conditions below are equivalent*:

1. *μ is absolutely continuous with respect to λ so that there exists a function $f(x)$ such that*

$$\mu(B) = \int_B f(x)d\lambda$$

2. *For all B, $\lambda(B) = 0$ implies $\mu(B) = 0$.*

The first condition asserts the existence of a "density function," as it is usually called in statistics, but it is the second condition that is usually referred to as absolute continuity. The function $f(x)$ is called the *Radon-Nikodym* derivative of μ with respect to λ. We sometimes write $f = \frac{d\mu}{d\lambda}$, but f is not in general unique. Indeed, there are many $f(x)$ corresponding to a single μ (i.e., many functions f satisfying $\mu(B) = \int_B f(x)d\lambda$ for all Borel B). However, for any two such functions f_1, f_2, $\lambda\{x; f_1(x) \neq f_2(x)\} = 0$. This means that f_1 and f_2 are *equal almost everywhere* (λ).

The so-called discrete distributions in statistics, such as the binomial distribution, the negative binomial, the geometric, the hypergeometric, the Poisson, or indeed any distribution concentrated on the integers, is absolutely continuous with respect to the counting measure $\lambda(A)$ =number of integers in A.

If the measure induced by a cumulative distribution function $F(x)$ is absolutely continuous with respect to Lebesgue measure, then $F(x)$ is a continuous function. However, it is possible that $F(x)$ is a continuous function without the corresponding measure being absolutely continuous with respect to Lebesgue measure.

Definition: Equivalent Measures Two measures μ and λ defined on the same measure space are said to be *equivalent* if both $\mu \ll \lambda$ and $\lambda \ll \mu$. Alternatively, they are equivalent if $\mu(A) = 0$ if and only if $\lambda(A) = 0$ for all A. Intuitively, this means that the two measures share exactly the same support or that the measures are either both positive on a given event or they are both zero on that event.

In general, there are three different types of probability distributions, when expressed in terms of the cumulative distribution function.

1. *Discrete*: For countable x_n, p_n, $F(x) = \sum_{\{n; x_n \leq x\}} p_n$. The corresponding measure has countably many atoms.
2. *Continuous singular*: $F(x)$ is a continuous function, but for some Borel set B having Lebesgue measure zero, $\lambda(B) = 0$, we have $P(X \in B) = \int_B dF(x) = 1$.
3. *Absolutely continuous* (with respect to Lebesgue measure): $F(x) = \int_{-\infty}^{x} f(x)d\lambda$ for some function f called the *probability density function*.

There is a general result called the Lebesgue decomposition, which asserts that any cumulative distribution function can be expressed as a mixture of those of the above three types; that is, a (sigma-finite) measure μ on the real line can be written

$$\mu = \mu_d + \mu_{ac} + \mu_s$$

the sum of a discrete measure μ_d, a measure μ_{ac} absolutely continuous with respect to Lebesgue measure, and a measure μ_s that is continuous singular. For a variety of reasons of dubious validity, statisticians concentrate on absolutely continuous and discrete distributions, excluding, as a general rule, those that are singular.

1.4 DISCRETE BIVARIATE AND MULTIVARIATE DISTRIBUTIONS

Definitions For discrete random variables X, Y defined on the same probability space, the function $f(x, y) = P[X = x, Y = y]$ giving the probability of all combinations of values of the random variables X, Y is called the *joint probability function* of X and Y. (Read the comma as the word "and," the intersection of two events.) The function $F(x, y) = P[X \leq x, Y \leq y]$ is called the *joint cumulative distribution function*. The joint probability function allows us to compute the probability functions of both X and Y. For example,

$$P[X = x] = \sum_{\text{all } y} f(x, y)$$

We call this the *marginal* probability function of X, denoted by $f_X(x) = P[X = x] = \sum_{\text{all } y} f(x, y)$. Similarly, $f_Y(y)$ is obtained by adding the joint probability function over all values of x. Finally, we are often interested in the conditional probabilities of the form

$$P[X = x|Y = y] = f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

This is called the *conditional probability function* of X given Y.

Expected Values For a single (discrete) random variable we determined the expected value of a function of X, say $h(X)$, by

$$E[h(X)] = \sum_{\text{all x}} (\text{value of } h) \times (\text{probability of value}) = \sum_x h(x)f(x)$$

For two or more random variables we should use a similar approach. However, when we add over all cases, this requires adding over all values of x and y. Thus, if h is a function of both X and Y,

$$E[h(X, Y)] = \sum_{\text{all } x \text{ and } y} h(x, y)f(x, y)$$

Independent Random Variables Two discrete random variables X, Y are said to be *independent* if the events $[X = x]$ and $[Y = y]$ are independent for all x, y, that is, if

$$P[X = x, Y = y] = P[X = x]P[Y = y] \quad \text{for all } x, y$$

or equivalently if

$$f(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y$$

This definition extends in a natural way to more than two random variables. For example, we say random variables $X_1, X_2, \ldots, X_n$ are (mutually) independent if, for every choice of values $x_1, x_2, \ldots, x_n$, the events $[X_1 = x_1], [X_2 = x_2], \ldots, [X_n = x_n]$ are independent events. This holds if the joint probability function of all n random variables factors into the product of the n marginal probability functions.

Theorem A6 *If X, Y are independent random variables, then*

$$E(XY) = E(X)E(Y)$$

Definition: Variance The variance of a random variable measures its variability about its own expected value. Thus if one random variable has larger variance than another, it *tends* to be farther from its own expectation. If we denote the expected value of X by $E(X) = \mu$, then

$$\mathrm{var}(X) = E[(X - \mu)^2]$$

Adding a constant to a random variable does not change its variance, but multiplying it by a constant does; it multiplies the original variance by the constant squared.

Example Suppose the random variable X has the binomial(n, p) distribution. Then

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \end{aligned}$$

$$\begin{aligned}
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1}(1-p)^{n-x} \\
&= np \left\{ \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-1-j)!} p^{j}\,(1-p)^{n-1-j} \right\} \\
&= np \left\{ \sum_{j=0}^{n-1} \binom{n-1}{j} p^{j}\,(1-p)^{n-1-j} \right\} \\
&= np
\end{aligned}$$

and so $E(X) = np$. A similar calculation allows us to obtain $E[X(X-1)] = n(n-1)p^2$, from which we can obtain $\text{var}(X) = np(1-p)$.

Definition: Covariance Define the covariance between 2 random variables X, Y as

$$\text{cov}(X, Y) = E[(X - EX)(Y - EY)]$$

Covariance measures the linear association between two random variables. Note that the covariance between two *independent random variables* is 0. If the covariance is large and positive, there is a tendency for large values of X to be associated with large values of Y. On the other hand, if large values of X are associated with small values of Y, the covariance will tend to be negative. There is an alternative form for covariance, generally easier for hand calculation but more subject to computer overflow problems: $\text{cov}(X, Y) = E(XY) - (EX)(EY)$.

Theorem A7 *For any two random variables* X, Y

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\,\text{cov}(X, Y)$$

One special case is of fundamental importance: the case when X, Y are independent random variables and $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$ since $\text{cov}(X, Y) = 0$.

Properties of Variance and Covariance For any random variables X_i and constants a_i

1. $\text{var}(X_1) = \text{cov}(X_1, X_1)$
2. $\text{var}(a_1X_1 + a_2) = a_1^2\,\text{var}(X_1)$
3. $\text{cov}(X_1, X_2) = \text{cov}(X_2, X_1)$
4. $\text{cov}(X_1, X_2 + X_3) = \text{cov}(X_1, X_2) + \text{cov}(X_1, X_3)$
5. $\text{cov}(a_1X_1, a_2X_2) = a_1a_2\,\text{cov}(X_1, X_2)$
6. Similarly, $\text{var}(\sum_{i=1}^{n} a_iX_i) = \sum a_i^2\,\text{var}(X_i) + 2\sum\sum_{\{(i,j);i<j\}} a_ia_j\,\text{cov}(X_i, X_j)$

Correlation Coefficient The covariance has an arbitrary scale factor because of property 5 above. This means that if we change the units in which something is measured (for example, a change from imperial to metric units of weight), the covariance will change. It is desirable to measure covariance in units free of the effect of scale. To this end, define the *standard deviation* of X by $SD(X) = \sqrt{\text{var}(X)}$. Then the *correlation coefficient* between X and Y is

$$\rho = \frac{\text{cov}(X, Y)}{SD(X)SD(Y)}$$

For any pair of random variables X, Y, we have $-1 \leq \rho \leq 1$ with $\rho = \pm 1$ if and only if the points (X, Y) always lie on a line, so $Y = aX + b$ (almost surely) for some constants a, b. The fact that $\rho \leq 1$ follows from the next argument, and the argument for $-1 \leq \rho$ is similar. Consider for any t,

$$\begin{aligned} \text{var}(X - tY) &= \text{cov}(X - tY, X - tY) \\ &= \text{var}(X) - 2t\ \text{cov}(X, Y) + t^2\ \text{var}(Y) \end{aligned}$$

Since variance is always ≥ 0, this quadratic equation in t cannot have two real roots, so the discriminant must be nonpositive,

$$[2\ \text{cov}(X, Y)]^2 - 4\ \text{var}(X)\ \text{var}(Y) \leq 0$$

that is,

$$|\text{cov}(X, Y)| \leq \sqrt{\text{var}(X)\ \text{var}(Y)}$$

The Multinomial Distribution Suppose an experiment is repeated n times (called "trials"), where n is fixed in advance. On each trial of the experiment, we obtain an outcome in one of k different categories $A_1, A_2, \ldots, A_k$, with the probability of outcome A_i given by p_i. Here $\sum_{i=1}^{k} p_i = 1$. At the end of the n trials of the experiment consider the count of $X_i =$ number of outcomes in category i, for $i = 1, 2, \ldots, k$. Then the random variables $(X_1, X_2, \ldots, X_k)$ have a joint *multinomial* distribution given by the joint probability function

$$\begin{aligned} P[X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k] &= \binom{n}{x_1 x_2 \cdots x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \\ &= \frac{n!}{x_1!x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \end{aligned}$$

whenever $\sum_i x_i = n$. Otherwise $P[X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k]$ is 0. Note that the marginal distribution of each X_i is binomial (n, p_i), and so $E(X_i) = np_i$.

Covariance of a Linear Transformation Suppose $X = (X_1, \ldots, X_n)'$ is a vector whose components are possibly dependent random variables. We define the expected value of this random vector by

$$\mu = E(X) = \begin{pmatrix} EX_1 \\ \vdots \\ EX_n \end{pmatrix}$$

and the covariance matrix by

$$V = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \ldots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1 & \mathrm{var}(X_2) & \ldots & \mathrm{cov}(X_2, X_n \\ \vdots & & & \vdots \\ \mathrm{cov}(X_n, X_1) & \ldots & \ldots & \mathrm{var}(X_n) \end{pmatrix}$$

Then if A is a $q \times n$ matrix of constants, the random vector $Y = AX$ has mean $A\mu$ and covariance matrix AVA'. In particular, if $q = 1$, the variance of AX is AVA' .

1.5 CONTINUOUS DISTRIBUTIONS

Definitions Suppose a random variable X can take any real number in an interval. Of course, the number that we record is often rounded to some appropriate number of decimal places, so we don't actually observe X but $Y = X$ rounded to the nearest $\Delta/2$ units. So, for example, the probability that we record the number $Y = y$ is the probability that X falls in the interval $y - \Delta/2 < X \leq y + \Delta/2$. If $F(x)$ is the cumulative distribution function of X, this probability is $P[Y = y] = F(y + \Delta/2) - F(y - \Delta/2)$. Suppose now that Δ is very small and that the cumulative distribution function is piecewise continuously differentiable with a derivative given in an interval by

$$f(x) = F'(x)$$

Then $F(y + \Delta/2) - F(y - \Delta/2) \approx f(y)\Delta$ and so Y is a discrete random variable with probability function given (approximately) by $P[Y = y] \approx \Delta f(y)$. The derivative of the cumulative distribution function of X is the *probability density function* of the random variable X. Notice that an interval of small length Δ around the point y has approximate probability given by *length of interval* $\times f(y)$. Thus the probability of a (small) interval is approximately proportional to the probability density function in that interval, and this is the motivation behind the term "probability density."

Example Suppose X is a random number chosen in the interval $[0, 1]$. We wish that any interval of length $\Delta \subset [0, 1]$ will have the same probability Δ regardless of where it is located. Then the cumulative distribution function is given by

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

The probability density function is given by the derivative of the cumulative distribution function $f(x) = 1$ for $0 < x < 1$ and $f(x) = 0$ otherwise. Notice that $F(y) = \int_{-\infty}^{y} f(x)dx$ for all y, and the probability density function can be used to determine probabilities as follows:

$$P[a < X < b] = P[a \le X \le b] = \int_a^b f(x)dx$$

In particular, notice that $F(b) = \int_{-\infty}^{b} f(x)dx$ for all b.

Example Let $F(x)$ be the binomial$(n, 1/2)$ cumulative distribution function. Notice that the derivative $F'(x)$ exists and is continuous (in fact is zero) except at finitely many points $x = 0, 1, 2, 3, 4$. Is it true that $F(b) = \int_{-\infty}^{b} F'(x)dx$? In this case the right side is zero since $F'(x) = 0$ except at finitely many points, but the left side is not. Equality is guaranteed only under further conditions.

Definition: Cumulative Distribution Function Suppose the cumulative distribution function of a random variable $F(x)$ is such that its derivative $f(x) = F'(x)$ exists except at finitely many points. Suppose also that

$$F(b) = \int_{-\infty}^{b} f(x)dx \tag{1.6}$$

for all $b \in \Re$. Then the distribution is *absolutely continuous* and the function $f(x)$ is called the *probability density function.*

Example Is it really necessary to impose the additional requirement (1.6), or this just a consequence of the fundamental theorem of calculus? Consider the case $F(x) = 0, x < 0$, and $F(x) = 1$, for $x \ge 0$. This cumulative distribution function is piecewise differentiable (the only point where the derivative fails to exist is the point $x = 0$). But is the function the integral of its derivative? Since the derivative is zero except at one point where it is not defined, any sensible notion of integral results in $\int_{-\infty}^{b} F'(x)dx = 0$ for any b.

For a continuous distribution, probabilities are determined by integrating the probability density function. Thus

$$P[a < X < b] = \int_a^b f(x)dx \tag{1.7}$$

A probability density function is not unique. For example, we may change $f(x)$ at finitely many points and it will still satisfy (1.7) above, and all probabilities, determined by integrating the function, remain unchanged. Whenever possible we will choose a continuous version of a probability density function; but at a finite number of discontinuity points, it does not matter how we define the function.

Properties of a Probability Density Function

1. $f(x) \geq 0$ for all $x \in \Re$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

The Continuous Uniform Distribution Consider a random variable X that takes values with a continuous uniform distribution on the interval $[a, b]$. Then the cumulative distribution function is

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

and so the probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$, and elsewhere the probability density function is 0. Again, notice that the definition of f at the points a and b does not matter, since altering the definition at two points will not alter the integral of the function.

Suppose we were to approximate a continuous random variable X having probability density function $f(x)$ by a discrete random variable Y obtained by rounding X to the nearest Δ units. Then the probability function of the discrete random variable Y is

$$P[Y = y] = P[y - \Delta/2 \leq X \leq y + \Delta/2] \approx \Delta f(y)$$

and its expected value is

$$E(Y) = \sum_y yP[y - \Delta/2 < X \leq y + \Delta/2] \approx \sum_y y\Delta f(y)$$

Note that as the interval length Δ approaches 0, this sum approaches the integral

$$\int xf(x)dx$$

This argues for the following definition of expected value for continuous random variables, if it is to be compatible with the expected value of its discretized or rounded relative Y. For *continuous random variables*

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

and for any function on the real numbers $h(x)$,

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

Using this definition, we find that for the uniform density $f(x) = \frac{1}{b-a}$ for $a < x < b$, the expected value is the midpoint between the two ends of the interval $\frac{a+b}{2}$.

The Exponential Distribution Consider a random variable X having probability density function

$$f(x) = \frac{1}{\mu}e^{-x/\mu}, \quad x > 0$$

The cumulative distribution function is given by

$$F(x) = 1 - e^{-x/\mu}$$

and the moments are

$$E(X) = \mu, \quad \mathrm{var}(X) = \mu^2$$

Such a random variable is called the *exponential distribution*, and it is commonly used to model lifetimes of simple components such as fuses and transistors.

The Normal Distribution

Normal Approximation to the Poisson Distribution Consider a random variable X that has the Poisson distribution with parameter μ. Recall that $E(X) = \mu$ and $\mathrm{var}(X) = \mu$, so $SD(X) = \mu^{1/2}$. We wish to approximate the distribution of this random variable for large values of μ. In order to prevent the distribution from disappearing off to $+\infty$, consider the standardized random variable

$$Z = \frac{X - \mu}{\mu^{1/2}}$$

Then $P[Z = z] = P[X = \mu + z\mu^{1/2}] = \frac{\mu^x}{x!}e^{-\mu}$, where $x = \mu + z\mu^{1/2}$ is an integer. Using Stirling's approximation $x! \sim \sqrt{2\pi x}x^x e^{-x}$and taking the limit of this as $\mu \to \infty$, we obtain

$$\frac{\mu^x}{x!}e^{-\mu} \sim \frac{1}{\sqrt{2\pi\mu}}e^{-z^2/2}$$

where the symbol $\sim$ is taken to mean that the ratio of the left- to the right-hand side approaches 1. The function on the right-hand side is a constant multiple of one of the basic functions in statistics, $e^{-x^2/2}$, which, upon normalization so that it integrates to one, is the standard normal probability density function.

The Standard Normal Distribution Consider a continuous random variable with probability density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad -\infty < x < \infty$$

Such a distribution we call the *standard normal distribution* or the $N(0, 1)$ distribution (Figure 1.1). The cumulative distribution function

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

is not obtainable in simple closed form, and requires either numerical approximation or a table of values. The probability density function $f(x)$ is symmetric about 0 and appears roughly as follows.

The integral of the standard normal probability density function is 1, but to show this requires conversion to polar coordinates. If we square the integral of the normal probability density function, we obtain

$$\begin{aligned}
&\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy\right)\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx\right)\\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}dx\,dy\\
&= \frac{1}{2\pi}\int_{0}^{\infty}\int_{0}^{2\pi} e^{-\frac{1}{2}r^2}r d\theta\,dr \quad \text{where } x = r\cos\theta \text{ and } y = r\sin\theta\\
&= 1
\end{aligned}$$

The normal cumulative distribution function is as given in Figure 1.2.

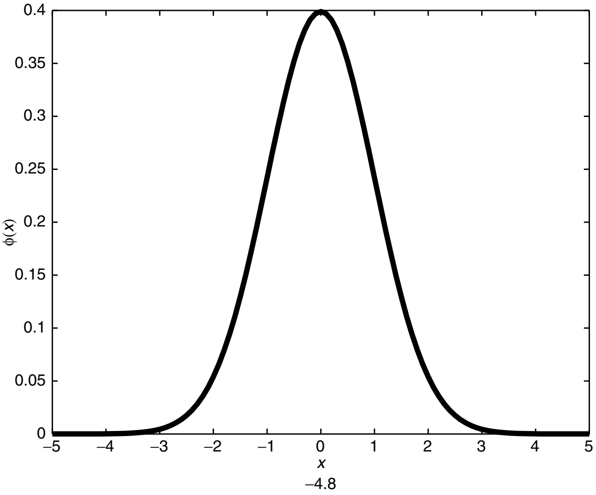

FIGURE 1.1 The Standard Normal Probability Density Function

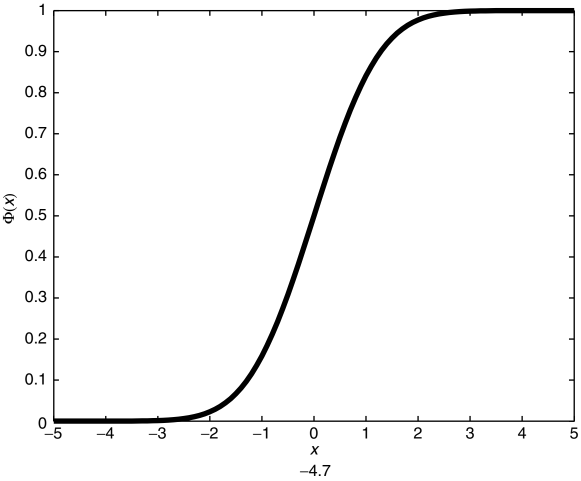

FIGURE 1.2 The Standard Normal Cumulative Distribution Function

TABLE 1.1 Values of the Standard Normal Cumulative Distribution Function $\Phi(x)$

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0.	.500	.540	.579	.618	.655	.692	.726	.758	.788	.816
1.	.841	.864	.885	.903	.919	.933	.945	.955	.964	.971
2.	.977	.982	.986	.989	.992	.994	.995	.997	.997	.998
3.	.9987	.9990	.9993	.9995	.9997	.9998	.9998	.9999	.9999	.99995

We usually provide the values of the normal cumulative distribution function either through a function such as *normcdf* in Matlab or through a table of values such as Table A1 (a much more compact version than that found at the back of most statistics books).

For example, we can obtain

$$\Phi(1.1) = 0.864 \qquad \Phi(0.6) = 0.726$$
$$\Phi(-0.5) = 1 - \Phi(0.5) = 1 - 0.692$$

Note, for example that $\Phi(-x) = 1 - \Phi(x)$ for all x, and if Z has a standard normal distribution, we can find probabilities of intervals such as

$$P[-1 < Z < 1] \approx 0.68 \quad \text{and} \quad P[-2 < Z < 2] \approx 0.95^4$$

The General Normal Distribution If we introduce a shift in the location in the graph of the normal density as well as a change in scale, then the resulting random variable is of the form

$$X = \mu + \sigma Z, \quad Z \sim N(0, 1)$$

for some constants $-\infty < \mu < \infty, \sigma > 0$. In this case, since

$$P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

it is easy to show by differentiating this with respect to x that the probability density function of X is

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

If a random variable X has the above normal distribution with location μ and scale σ, we will denote this by $X \sim N(\mu, \sigma^2)$.

Moments Show that the function $f(x;\mu,\sigma)$ integrates to 1 and is therefore a probability density function. It is not too hard to find the expected value and variance of a random variable having the probability density function $f(x;\mu,\sigma)$ by integration:

$$E(X) = \int_{-\infty}^{\infty} xf(x;\mu,\sigma)dx = \mu$$
$$\mathrm{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x;\mu,\sigma)dx = \sigma^2$$

and this gives meaning to the parameters μ and σ^2, the former being the mean or expected value of the distribution and the latter the variance.

Linear Combinations of Normal Random Variables Suppose $X_1 \sim N(\mu_1,\sigma_1^2)$ and $X_2 \sim N(\mu_2,\sigma_2^2)$ are independent random variables. Then $X_1+X_2 \sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$. More generally if we sum independent random variables, each having a normal distribution, the sum itself also has a normal distribution. The expected value of the sum is the sum of the expected values of the individual random variables, and the variance of the sum is the sum of the variances.

Problem Suppose $X_i \sim N(\mu,\sigma^2)$ are independent random variables. What is the distribution of the sample mean

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

Assume $\sigma = 1$ and find the probability $P[|\overline{X}_n - \mu| > 0.1]$ for various values of n. What happens to this probability as $n \to \infty$?

The Central Limit Theorem

The major reason that the normal distribution is the single most commonly used distribution is the fact that it tends to approximate the distribution of sums of random variables. For example, if we throw n dice and S_n is the sum of the outcomes, what is the distribution of S_n? The tables below provide the number of ways in which a given value can be obtained. The corresponding probability is obtained by dividing by 6^n. For example, on the throw of $n = 1$ die the probable outcomes are $1, 2, \ldots, 6$ with probabilities all 1/6 as indicated in Figure 1.3.

If we sum the values on two fair dice, the possible outcomes are the values 2,3,...,12 as shown in the following table and the probabilities are

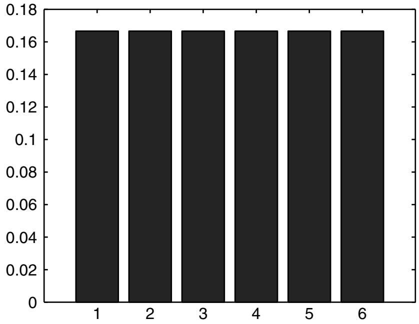

FIGURE 1.3 The Sum of $n = 1$ Discrete Uniform $\{1, 2, 3, 4, 5, 6\}$ Random Variables

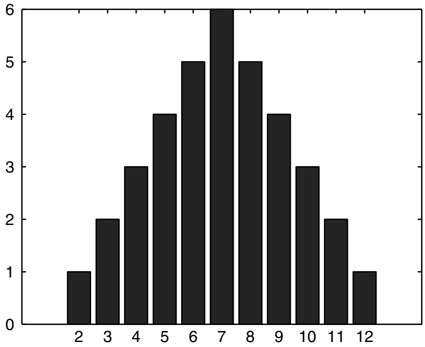

FIGURE 1.4 The Sum of $n = 2$ Discrete Uniform $\{1, 2, 3, 4, 5, 6\}$ Random Variables

the values below:

Values	2	3	4	5	6	7	8	9	10	11	12
Probabilities × 36	1	2	3	4	5	6	5	4	3	2	1

The probability histogram of these values is shown in Figure 1.4.

Finally, for the sum of the values on three independent dice, the values range from 3 to 18 and have probabilities which, when multiplied by 6^3, result in the values

1	3	6	10	15	21	25	27	27	25	21	15	10	6	3	1

to which we can fit three separate quadratic functions—one in the middle region and one in each of the two tails. The histogram of these values in Figure 1.5 already resembles a normal probability density function.

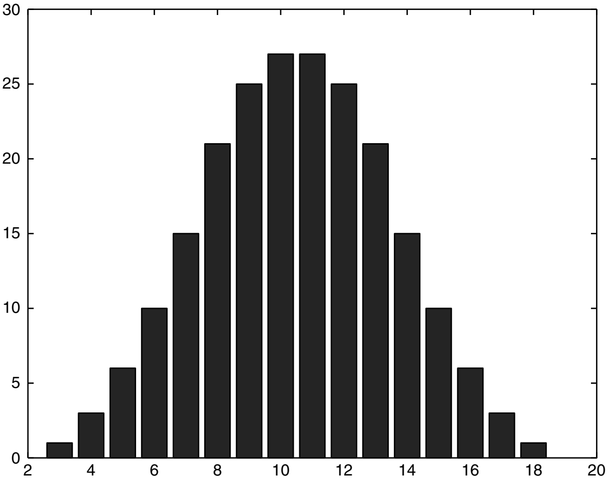

FIGURE 1.5 The Distribution of the Sum of Three Discrete Uniform {1, 2, 3, 4, 5, 6} Random Variables

In general, these distributions show a simple pattern. For $n = 1$, the probability function is a constant (polynomial degree 0). For $n = 2$, it is two linear functions spliced together. For $n = 3$, it is a spline consisting of three quadratic pieces (polynomials of degree $n - 1$). In general, the histogram for S_n, the sum of the values on n independent dice, consists of n piecewise polynomials of degree $n - 1$. These histograms rapidly approach the shape of the normal probability density function.

Example Let $X_i = 0$ or 1 when the ith toss of a biased coin is tails or heads, respectively. What is the distribution of $S_n = \sum_{i=1}^{n} X_i$? Consider the standardized random variable obtained by subtracting $E(S_n)$ and dividing by its standard deviation or the square root of var(S_n):

$$S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}}$$

Suppose we approximate the distribution of S_n^* for large values of n.

First, consider a sequence of integers $x = x_n$ that are close to the real number $np + \sqrt{np(1-p)}$ in the sense that the difference is bounded by a constant. Mathematically we write $x \sim np + z\sqrt{np(1-p)}$ for fixed z and $0 < p < 1$. Then as $n \to \infty$, $x/n \to p$. Stirling's approximation tells us that $n! \sim \sqrt{2\pi} n^{n+1/2} e$ so that

$$\binom{n}{x} \sim \frac{\sqrt{2\pi} n^{n+1/2} e^{-n}}{2\pi x^{x+1/2} (n-x)^{n-x+1/2}} \sim \frac{1}{\sqrt{2\pi np(1-p)}\,(\frac{x}{n})^{x}(1-\frac{x}{n})^{n-x}}$$

Also using the series expansion $\ln(1+x) = x - \frac{1}{2}x^2 + O\left(x^3\right)$, setting $\sigma = \sqrt{\frac{p(1-p)}{n}}$, and noting that $\sigma \to 0$ as $n \to \infty$,

$$
\begin{aligned}
&\ln\left\{\frac{p^x(1-p)^{n-x}}{(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}}\right\} \\
&\quad = x\ln\left(\frac{p}{p+z\sigma}\right) + (n-x)\ln\left(\frac{1-p}{1-p-z\sigma}\right) \\
&\quad = -x\ln\left(1+\frac{z\sigma}{p}\right) - (n-x)\ln\left(1-\frac{z\sigma}{1-p}\right) \\
&\quad = -n(p+z\sigma)\ln\left(1+\frac{z\sigma}{p}\right) - n(1-p-z\sigma)\ln\left(1-\frac{z\sigma}{1-p}\right) \\
&\quad = -n(p+z\sigma)\left\{\left(\frac{z\sigma}{p}\right) - \frac{1}{2}\left(\frac{z\sigma}{p}\right)^2 + O\left(\frac{z\sigma}{p}\right)^3\right\} \\
&\qquad - n(1-p-z\sigma)\left\{-\left(\frac{z\sigma}{1-p}\right) - \frac{1}{2}\left(\frac{z\sigma}{1-p}\right)^2 + O\left(\frac{z\sigma}{1-p}\right)^3\right\} \\
&\quad = -n\left\{z\sigma + \frac{z^2\sigma^2}{p} - \frac{1}{2}\frac{z^2\sigma^2}{p} - z\sigma + \frac{z^2\sigma^2}{1-p} - \frac{1}{2}\frac{z^2\sigma^2}{1-p} + O(\sigma^3)\right\} \\
&\quad = -\frac{1}{2}z^2\sigma^2\left(\frac{n}{p} + \frac{n}{1-p}\right) + O(n^{-1/2}) = -\frac{z^2}{2} + O(n^{-1/2})
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
P[S_n = x] = P[S_n^* = z] &= \binom{n}{x}p^x(1-p)^{n-x} \\
&\sim \binom{n}{x}\left(\frac{x}{n}\right)^x\left(1-\frac{x}{n}\right)^{n-x}\frac{p^x(1-p)^{n-x}}{(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}} \\
&\sim \frac{1}{\sqrt{np(1-p)}}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}
\end{aligned}
$$

This is the standard normal probability density function multiplied by the distance between consecutive values of S_n^*. In other words, this result says that the area under the probability histogram for S_n^* for the bar around the point z can be approximated by the area under the normal curve between the same two points $\left(z \pm \frac{1}{2\sqrt{np(1-p)}}\right)$.

Theorem A8 *Let $X_i, i = 1, \ldots, n$, be independent random variables all with the same distribution, and with mean μ and variance σ^2. Then the cumulative*

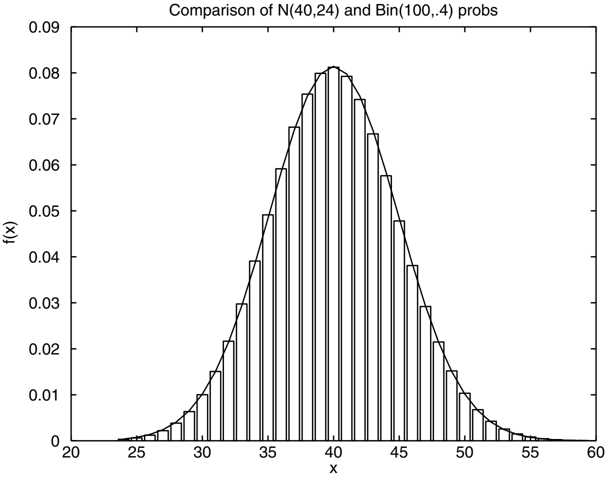

FIGURE 1.6 Binomial(100, 0.4) Probability Histogram Together with $N(40, 24)$ Probability Density Function

distribution function of

$$S_n^* = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

converges to the cumulative distribution function of a standard normal random variable.

Consider, for example, the case where the X_i are independent, each with a Bernoulli(p) distribution. Then the sum $\sum_{i=1}^{n} X_i$ has a binomial distribution with parameters n, p and the above theorem asserts that if we subtract the mean and divide by the standard deviation of a binomial random variable, the result is approximately standard normal. In other words, for large values of n a binomial random variable is approximately normal($np, np(1-p)$). To verify this fact, we plot both the binomial(100, 0.4) histogram as well as the normal probability density function in Figure 1.6.

Problem Use the central limit theorem and the normal approximation to a probability histogram to estimate the probability that the sum of the numbers on 5 dice is 15. Compare your answer with the exact probability.

The Distribution of a Function of a Random Variable We have seen that if X has a normal distribution, then a linear function of X, say $aX + b$, also has a normal distribution. The parameters are easily determined since $E(aX+b) = aE(X)+b$ and $\text{var}(aX+b) = a^2\,\text{var}(X)$. Is this true of arbitrary functions and general distributions? For example, is X^2 normally distributed? The answer in general is *no*. For example, the distribution of X^2 must be concentrated entirely on the positive values of x, whereas the normal distributions are all supported on the whole real line (i.e., the probability density function $f(x) > 0$, all $x \in \mathcal{R}$). In general, the safest method for finding the distribution of the function of a random variable in the continuous case is to first find the cumulative distribution of the function and then differentiate to obtain the probability density function. This allows us to verify the result below.

Theorem A9 *Suppose a continuous random variable X has probability density function $f_X(x)$. Then the probability density function of $Y = h(X)$ where $h(.)$ is a continuous monotone increasing function with inverse function $h^{-1}(y)$ is*

$$f_Y(y) = f_X(h^{-1}(y))\frac{d}{dy}h^{-1}(y)$$

1.6 MOMENT-GENERATING FUNCTIONS

Consider a random variable X. We have seen several ways of describing its distribution, using either a cumulative distribution function, a probability density function (continuous case) or probability function, or a probability histogram or table (discrete case). We may also use some transform of the probability density or probability function. For example, consider the function

$$M_X(t) = Ee^{tX}$$

defined for all values of t such that this expectation exists and is finite. This function is called the moment-generating function of the (distribution of the) random variable X. It is a powerful tool for determining the distribution of sums of independent random variables and for proving the central limit theorem. In the discrete case we can write $M_X(t) = \sum_x e^{xt}P[X = x]$, and in the continuous case $M_X(t) = \int_{-\infty}^{\infty} e^{xt}f(x)dx$. The logarithm of the moment-generating function $\ln(M_X(t))$ is called the cumulant-generating function.

Properties of the Moment-Generating Function For these properties we assume that the moment-generating function exists at least in some neighborhood of the value $t = 0$, say for $-\epsilon < t < \epsilon$ for some $\epsilon > 0$. We also assume that $\frac{d}{dt}E[X^n e^{tX}] = E[\frac{d}{dt}X^n e^{tX}]$ for each value of $n = 0, 1, 2, \ldots$ then for $-\epsilon < t < \epsilon$. The ability to differentiate under an integral or infinite sum is justified under general conditions involving the rate at which the integral or series converges.

1. $M'(0) = E(X)$.
2. $M^{(n)}(0) = E(X^n), n = 1, 2, \ldots$.
3. A moment-generating function uniquely determines a distribution. In other words, if $M_X(t) = M_Y(t)$ for all $-\epsilon < t < \epsilon$, then X and Y have the same distribution.
4. $M_{aX+b}(t) = e^{bt}M_X(at)$ for constants a, b.
5. If X and Y are independent random variables, $M_{X+Y}(t) = M_X(t)M_Y(t)$.

Examples Let X have a distribution as given in the first column of the table below. Then the moment-generating function of X is as given in column 2.

Distribution	Moment-Generating Function $M_X(t)$
Binomial(n, p)	$(pe^t + 1 - p)^n$
Poisson$(\lambda\ )$	$\exp\{\lambda(e^t - 1)\}$
Exponential, mean μ	$\frac{1}{1-\mu t}$ for $t < 1/\mu$
Normal(μ, σ^2)	$\exp\{\mu t + \sigma^2 t^2/2\}$

Moment-generating functions are useful for showing that a sequence of cumulative distribution functions converge because of the following result. The result implies that convergence of the moment-generating functions can be used to show convergence of the cumulative distribution functions (i.e., convergence of the distributions).

Theorem A10 *Suppose Z_n is a sequence of random variables with moment-generating functions $M_n(t)$. Let Z be a random variable Z having moment-generating function $M(t)$. If $M_n(t) \to M(t)$ for all t in a neighborhood of 0, then*

$$P[Z_n \le z] \to P[Z \le z]$$

as $n \to \infty$ for all values of z at which the function $F_Z(z)$ is continuous.

1.7 JOINT DISTRIBUTIONS AND CONVERGENCE

Consider constructing measures on a product Euclidean space. Given Lebesgue measure λ, essentially a measure of length on the real line $\Re$, how do we construct a similar measure compatible with the notion of area in two-dimensional Euclidean space? We naturally begin with the measure of rectangles or indeed any *product* set of the form $A \times B = \{(x, y); x \in A,\ y \in B\}$ for arbitrary (Borel) sets $A \subset \Re, B \subset \Re$. The measure of a product set can be defined as the product of the measure of the two-factor sets $\mu(A \times B) = \lambda(A)$ $\lambda(B)$. This defines a measure for any product set, and by an extension theorem, since the product sets form a Boolean algebra, we can extend this measure to the sigma algebra generated by the product sets.

More formally, suppose we are given two measure spaces $(M, \mathcal{M}, \mu)$ and $(N, \mathcal{N}, \nu)$. Define the *product space* to be the space consisting of pairs of objects, one from each of M and N,

$$\Omega = M \times N = \{(x, y);\ x \in M,\ y \in N\}$$

The Cartesian product of two sets $A \subset M,\ \ B \subset N$ is denoted $A \times B = \{(a, b); a \in A, b \in B\}$. This is the analogue of a rectangle, a subset of $M \times N$, and it is easy to define a measure for such sets as follows. Define the *product measure* of product sets of the above form by $\pi(A \times B) = \mu(A)\nu(B)$. The following theorem is a simple consequence of the Caratheodory extension theorem.

Theorem A11 *The product measure π defined on the product sets of the form $\{A \times B; A \in \mathcal{N},\ \mathcal{B} \in \mathcal{M}\}$ can be extended to a measure on the sigma algebra $\sigma\{A \times B; A \in \mathcal{N}, \mathcal{B} \in \mathcal{M}\}$ of subsets of $M \times N$.*

There are two cases of product measure of importance. Consider the sigma algebra on $\Re^2$ generated by the product of the Borel sigma algebras on $\Re$. This is called the Borel sigma algebra in $\Re^2$. We can similarly define the Borel sigma algebra on $\Re^n$.

Similarly, if we are given two probability spaces $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$, we can construct a *product measure* Q on the Cartesian product space $\Omega_1 \times \Omega_2$ such that $Q(A \times B) = P_1(A)P_2(B)$ for all $A \in \mathcal{F}_1,\ B \in \mathcal{F}_2$. This guarantees the existence of a product probability space in which events of the form $A \times \Omega_2$ are independent of events of the form $\Omega_1 \times B$ for $A \in \mathcal{F}_1, B \in \mathcal{F}_2$.

We say a sequence of random variables $X_1, X_2, \ldots$ is *independent* if the family of sigma algebras $\sigma(X_1), \sigma(X_2), \ldots$ are independent; that is, for Borel sets $B_n, n = 1, \ldots, N$ in $\Re$, the events $[X_n \in B_n], n = 1, \ldots, N$ form a mutually independent sequence of events so that

$$P[X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n] = P[X_1 \in B_1]P[X_2 \in B_2] \cdots P[X_n \in B_n]$$

The sequence is said to be *identically distributed* if every random variable X_n has the same cumulative distribution function.

We have already seen the following result, but we repeat it here, if only to get the flavor of the proof.

If X, Y are independent integrable random variables on the same probability space, then XY is also integrable and

$$E(XY) = E(X)E(Y)$$

Proof. Suppose first that X and Y are both simple functions, $X = \sum c_i I_{A_i}$, $Y = \sum d_j I_{B_j}$. Then X and Y are independent if and only if $P(A_i B_j) = P(A_i)P(B_j)$ for all i, j, and so

$$\begin{aligned} E(XY) &= E[(\sum c_i I_{A_i})(\sum d_j I_{B_j})] \\ &= \sum\sum c_i d_j E(I_{A_i} I_{B_j}) \\ &= \sum\sum c_i d_j P(A_i)P(B_j) \\ &= E(X)E(Y) \end{aligned}$$

More generally, suppose X, Y are nonnegative random variables and consider independent simple functions X_n increasing to X and Y_n increasing to Y. Then $X_n Y_n$ is a nondecreasing sequence with limit XY. Therefore, by the monotone convergence theorem,

$$E(X_n Y_n) \to E(XY)$$

On the other hand,

$$E(X_n Y_n) = E(X_n)E(Y_n) \to E(X)E(Y).$$

Therefore, $E(XY) = E(X)E(Y)$. The case of general (positive and negative random variables X, Y we leave as a problem. ■

Joint Distributions of More Than Two Random Variables Suppose $X_1, \ldots, X_n$ are random variables defined on the same probability space $(\Omega, \mathcal{F}, P)$ (but not necessarily independent). The joint distribution can be characterized by the *joint cumulative distribution function*, a function on $\Re^n$ defined by

$$F(x_1, \ldots, x_n) = P[X_1 \le x_1, \ldots, X_n \le x_n] = P([X_1 \le x_1] \cap \cdots \cap [X_n \le x_n])$$

The joint cumulative distribution function allows us to find $P[a_1 < X_1 \le b_1, \ldots, a_n < X_n \le b_n]$. By the inclusion-exclusion principle,

$$\begin{aligned} P[a_1 < X_1 \le b_1, \ldots, a_n < X_n \le b_n] &= F(b_1, b_2, \ldots, b_n) \\ &- \sum_j F(b_1, \ldots, a_j, b_{j+1}, \ldots, b_n) \\ &+ \sum_{i<j} F(b_1, \ldots, a_i, b_{i+1}, \ldots, a_j, b_{j+1}, \ldots, b_n) - \cdots \qquad (1.8) \end{aligned}$$

The formula (1.8) above allows us to construct the probability measure of any product of intervals

$$C = (a_1, b_1] \times (a_2, b_2] \times \cdots (a_n, b_n]$$

and thereby any disjoint union of finitely many sets of the form C. The class of all such disjoint unions (including all of $\Re^n$) forms an algebra of sets, closed under complements, finite unions and intersections. In the same way as we constructed Lebesgue measure on the Euclidean space $\Re^n$ from the basic notion of the length of an interval, we can now extend this probability measure to all sets in the sigma algebra generated by sets C of the form above. In general, a joint cumulative distribution function defined on $\Re^n$ allows us to define a probability measure on $n-$dimensional Euclidean space. However in order that a function qualify as a joint c.d.f., the following conditions need to be satisfied.

Theorem A12 *The joint cumulative distribution function has the following properties:*

(a) $F(x_1, \ldots, x_n)$ *is right-continuous and nondecreasing in each argument* x_i *when the other arguments* x_j, $j \neq i$, *are fixed.*

(b) $F(x_1, \ldots, x_n) \to 1$ *as* $\min(x_1, \ldots, x_n) \to \infty$ *and* $F(x_1, \ldots, x_n) \to 0$ *as* $\min(x_1, \ldots, x_n) \to -\infty$.

(c) *The expression on the right-hand side of* (1.8) *is greater than or equal to zero for all* $a_1 < b_1, a_2 < b_2, \ldots, a_n < b_n$.

The joint probability distribution of the variables $X_1, \ldots, X_n$ is a measure on $\mathcal{R}^n$. It can be determined from the cumulative distribution function in the usual fashion, first by defining the measure of intervals and then extending this to the sigma algebra generated by these intervals. In order to verify that the random variables are mutually independent, it is sufficient to verify that the joint cumulative distribution function factors

$$F(x_1, \ldots, x_n) = F_1(x_1)F_2(x_2)\cdots F_n(x_n) = P[X_1 \leq x_1]\cdots P[X_n \leq x_n]$$

for all $x_1, \ldots, x_n \in \Re$.

Theorem A13 *If the random variables* $X_1, \ldots, X_n$ *are mutually independent, then*

$$E[\prod_{j=1}^{n} g_j(X_j)] = \prod_{j=1}^{n} E[g_j(X_j)]$$

for any Borel measurable functions $g_1, \ldots, g_n$.

An infinite sequence of random variables $X_1, X_2, \ldots$ is mutually independent if every finite subset is mutually independent.

Definition: Strong (Almost Sure) Convergence Let X and $X_n, n = 1, 2, \ldots,$ be random variables all defined on the same probability space $(\Omega, \mathcal{F})$. We say that the sequence X_n converges *almost surely* (or *with probability* 1) to X (denoted $X_n \to X$ *a.s.*) if the event

$$\{\omega; X_n(\omega) \to X(\omega)\} = \cap_{m=1}^{\infty} \left[|X_n - X| \leq \frac{1}{m} \ a.b.f.o. \right]$$

has probability 1. Here the notation *a.b.f.o.*, standing for "all but finitely often," is the "lim inf" of the events $[|X_n - X| \leq \frac{1}{m}]$.

In order to show that a sequence X_n converges almost surely, we need that X_n are (measurable) random variables for all n, and to show that there is some measurable random variable X for which the set $\{\omega; X_n(\omega) \to X(\omega)\}$ is measurable and hence an event, and that the probability of this event $P[X_n \to X]$ is 1. Alternatively, we can show that for each value of $\epsilon > 0$, $P[|X_n - X| > \epsilon \ i.o.] = 0$, or in other words, that the probability of the set of all points ω such that $X_n(\omega)$ does not converge to $X(\omega)$ is zero. It is sufficient to consider values of ϵ of the form $\epsilon = 1/m$, $m = 1, 2, \ldots$ above.

The law of large numbers (sometimes called the law of averages) is the best-known result in probability. It says, for example, that the average of independent Bernoulli random variables, or Poisson, or negative binomial, or gamma random variables, to name a few, converge to their expected value **with probability 1**.

Theorem A14 (Strong Law of Large Numbers) *If $X_n, n = 1, 2, \ldots,$ is a sequence of independent identically distributed random variables with $E|X_n| < \infty$ (i.e., they are integrable) and $E(X_n) = \mu$, then*

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to \mu \text{ almost surely as } n \to \infty$$

1.8 WEAK CONVERGENCE (CONVERGENCE IN DISTRIBUTION)

Consider random variables that are constants: $X_n(w) = 1 + \frac{1}{n}$ for all w. By any sensible definition of convergence, X_n converges to $X = 1$ as $n \to \infty$. Does the cumulative distribution function of X_n, say F_n, converge to the cumulative distribution function of X pointwise? In this case it is true that $F_n(x) \to F(x)$ at all values of x except the value $x = 1$, where the function $F(x)$ has a discontinuity. Convergence in distribution (weak convergence,

convergence in law) is defined as pointwise convergence of the c.d.f. at all values of x except those at which $F(x)$ is discontinuous. Of course, if the limiting distribution is absolutely continuous (for example, the normal distribution as in the Central Limit Theorem), then $F_n(x) \to F(x)$ does hold for all values of x.

Definition: Weak Convergence If $F_n(x)$, $n = 1, \ldots$, is a sequence of cumulative distribution functions and if F is a cumulative distribution function, we say that F_n converges to F *weakly* or *in distribution* if $F_n(x) \to F(x)$ for all x at which $F(x)$ is continuous. Weak convergence of a sequence of random variables X_n whose c.d.f. converges in the above sense is denoted in a variety of ways, such as $X_n \Rightarrow X$ or $X_n \to_D X$ (here D stands for "in distribution").

There are simple examples of cumulative distribution functions that converge pointwise but not to a genuine cumulative distribution because some of the mass of the distribution escapes to infinity. For example, if F_n is the cumulative distribution function of a point mass at the point n, then $F_n(x) \to 0$ for each fixed value of $x < \infty$. An additional condition, called tightness, is needed to ensure that the limiting distribution is a "proper" probability distribution (i.e., has total measure 1). A sequence of probability measures P_n on Euclidean space is *tight* if for all $\epsilon > 0$, there exists a bounded rectangle K such that $P_n(K) > 1 - \epsilon$ for all n. A sequence of cumulative distribution functions F_n defined on $\mathcal{R}$ is tight if, for every $\epsilon > 0$, there is a real number $M < \infty$ such that the probabilities of interval $[-M, M]$ are greater than than $1 - \epsilon$,

$$F_n(M) - F_n(-M) \leq 1 - \epsilon \quad \text{for all } n = 1, 2, \ldots$$

Tightness is a condition that ensures that none of the probability mass escapes to infinity. For example suppose a sequence of cumulative distribution functions $F_n(x)$ converges to some limiting right-continuous function $F(x)$ at all continuity points x of F and suppose the sequence F_n is tight. Then it is easy to show that the limiting function F is a proper cumulative distribution function (i.e. has total mass 1) and the convergence is in distribution.

There is an alternative definition of weak convergence that is more appropriate for more general spaces of random elements such as spaces of continuous time stochastic processes.

General Definition of Weak Convergence A sequence of random elements of a metric space X_n converges weakly to X (i.e., $X_n \Rightarrow X$) if and only if $E[f(X_n)] \to E[f(X)]$ for all bounded continuous functions f.

Definition: Convergence in Probability We say a sequence of random variables $X_n \to X$ *in probability* if for all $\epsilon > 0$, $P[|X_n - X| > \epsilon] \to 0$ as $n \to \infty$.

Convergence in probability is in general a somewhat more demanding concept than weak convergence, but less demanding than almost sure convergence. In other words, convergence almost surely implies convergence in probability, and convergence in probability implies weak convergence.

Theorem A15 *If* $X_n \to X$ almost surely, *then* $X_n \to X$ in probability.

However, convergence in probability does not imply convergence almost surely, but it does imply weak convergence.

Theorem A16 *If* $X_n \to X$ in probability, *then* $X_n \to_D X$.

The converse of this theorem holds under one condition, when the convergence in distribution s to a constant.

Theorem A17 *If* $X_n \to_D$ converges in distribution *to some constant c, then* $X_n \to c$ *in probability.*

The next result, Fubini's theorem, allows us to change the order of integration as long as the function being integrated is, in fact, integrable.

Theorem A18 (Fubini's Theorem) *Suppose* $g(x, y)$ *is integrable with respect to a product measure* $\pi = \mu \times \nu$ *on* $M \times N$. *Then*

$$\int_{M\times N} g(x, y) d\pi = \int_M \left[\int_N g(x, y) d\nu\right] d\mu = \int_N \left[\int_M g(x, y) d\mu\right] d\nu$$

Convolutions Consider two independent random variables X, Y, both having a discrete distribution. Suppose we wish to find the probability function of the sum $Z = X + Y$. Then

$$P[Z = z] = \sum_x P[X = x]P[Y = z - x] = \sum_x f_X(x) f_Y(z - x)$$

Similarly, if X, Y are independent, absolutely continuous distributions with probability density functions f_X, f_Y, respectively, then we find the probability density function of the sum $Z = X + Y$ by

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

In both the discrete and continuous cases, we can rewrite the above in terms of the cumulative distribution function F_Z of Z. In either case,

$$F_Z(z) = E[F_Y(z - X)] = \int_{\Re} F_Y(z - x)F_X(dx)$$

We use the last form as a more general definition of a *convolution* between two cumulative distribution functions F, G. We define the *convolution* of F and G to be $F * G(x) = \int_{-\infty}^{\infty} F(x - y)dG(y)$.

Properties of Convolution

(a) If F, G are cumulative distributions functions, then so is $F * G$.
(b) $F * G = G * F$.
(c) If either F or G is absolutely continuous with respect to Lebesgue measure, then $F * G$ is absolutely continuous with respect to Lebesgue measure.

The convolution of two cumulative distribution functions $F * G$ represents the c.d.f of the sum of two independent random variables, one with c.d.f. F and the other with c.d.f. G.

1.9 STOCHASTIC PROCESSES

A stochastic process is an indexed family of random variables X_t for t ranging over some index set T, such as the integers or an interval of the real line. For example, a sequence of independent random variables is a stochastic process, as is a Markov chain. For an example of a continuous-time stochastic process, define X_t to be the price of a stock at time t (assuming trading occurs continuously over time).

Markov Chains Consider a sequence of (discrete) random variables $X_1, X_2, \ldots,$ each of which takes integer values $1, 2, \ldots, N$ (called *states*). We assume that for a certain matrix P (called the *transition probability matrix*), the conditional probabilities are given by corresponding elements of the matrix,

$$P[X_{n+1} = j|X_n = i] = P_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, N$$

and furthermore that the chain cares only about the last state occupied in determining its future:

$$\begin{aligned} &P\left[X_{n+1} = j|X_n = i, X_{n-1} = i_1, X_{n-2} = i_2 \cdots X_{n-l} = i_l\right] \\ &\quad = P[X_{n+1} = j|X_n = i] = P_{ij} \end{aligned}$$

for all $j, i, i_1, i_2, \ldots, i_l$, and $l = 2, 3, \ldots$. Then the sequence of random variables X_n is called a *Markov chain*. Markov chain models are the most common simple models for dependent variables, including weather (precipitation, temperature), movements of security prices, and others. They allow the future of the process to depend on the present state of the process, but the past behavior can influence the future only through the present.

Properties of the Transition Matrix P Note that $P_{ij} \geq 0$ for all i, j and $\sum_j P_{ij} = 1$ for all i. This last property implies that the $N \times N$ matrix $P - I$ (where I is the identity matrix) has rank at most $N - 1$ because the sum of the N columns of $P - I$ is identically 0.

Example: Rain/No Rain Suppose that the probability that tomorrow is rainy given that today is not is α, and the probability that tomorrow is dry given that today is rainy is β. Then if we assume that tomorrow's weather depends on the past only through whether today is wet or dry, the chain given by

$$X_n = \begin{cases} 1 & \text{Day } n \text{ is wet} \\ 0 & \text{Day } n \text{ is dry} \end{cases}$$

is a Markov chain having transition matrix

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

The Distribution of X_n Suppose that the chain is started by randomly choosing a state for X_0 with distribution $P[X_0 = i] = q_i, i = 1, 2, \ldots, N$. Then the distribution of X_1 is given by

$$\begin{aligned} P(X_1 = j) &= \sum_{i=1}^{N} P(X_1 = j, X_0 = i) \\ &= \sum_{i=1}^{N} P(X_1 = j | X_0 = i) P(X_0 = i) \\ &= \sum_{i=1}^{N} P_{ij}\, q_i \end{aligned}$$

and this is the jth element of the vector $\underline{q}'P$, where $\underline{q}$ is the column vector of values q_i. Similarly the distribution of X_n is the vector $\underline{q}'P^n$, where P^n is the product of the matrix P with itself n times. Under very general conditions, it can be shown that these probabilities converge, and in many such cases the limit does not depend on the initial distribution $\underline{q}$.

Definition A *limiting distribution* of a Markov chain is a vector ($\underline{\pi}$, say) of long-run probabilities of the individual states so that

$$\pi_i = \lim_{t\to\infty} P[X_t = i]$$

Definition A *stationary distribution* of a Markov chain is the column vector ($\underline{\pi}$, say) of probabilities of the individual states such that

$$\underline{\pi}'P = \underline{\pi}'$$

Theorem A19 *Any limiting distribution of a Markov chain must be a stationary distribution.*

Proof. Note that $\pi' = \lim_{n\to\infty} \underline{q}'P^n = \lim_{n\to\infty}(\underline{q}'P^n)P = (\lim_{n\to\infty} \underline{q}'P^n)P = \underline{\pi}'P$. ■

Example: Binary Information Suppose that $X_1, X_2, \ldots$ is a sequence of binary information (Bernoulli random variables) taking values either 0 or 1. Suppose that the probability that a 0 is followed by a 1 is p and the probability that a 1 is followed by a 0 is given by q, where $0 < p, q < 1$. Then the transition matrix for the Markov chain is

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

The limiting distribution for this Markov chain is

$$\underline{\pi} = \begin{pmatrix} \frac{q}{p+q} \\ \frac{p}{p+q} \end{pmatrix}$$

So, for example, the long-run proportion of zeros in the sequence is $\frac{q}{p+q}$.

When is the limiting distribution of a Markov chain unique and independent of the initial state of the chain?

Definition: Irreducible, Aperiodic We say that a Markov chain is *irreducible* if every state can be reached from every other state. In other words, for every pair i, j there is some m such that $P_{i,j}^{(m)} > 0$. We say that the chain is *aperiodic* if, for each state i, there is no regular or periodic pattern for the values of k for which $P_{ii}^{(k)} > 0$. For example, if $P_{ii}^{(1)} = 0,\ P_{ii}^{(2)} > 0,\ P_{ii}^{(3)} = 0,\ P_{ii}^{(4)} > 0$ and this pattern continues indefinitely, then the greatest common divisor of the values k such that $P_{ii}^{(N)} > 0$ is evidently 2. We write this mathematically as $\gcd\{k; P_{ii}^{(k)} > 0\} = 2$, and this chain is not aperiodic; it has period 2.

On the other hand, if for all states i, $\gcd\{k; P_{ii}^{(k)} > 0\} = 1$, we say the chain is aperiodic. For a *periodic chain* (i.e., one that is not aperiodic, so the period $\gcd\{k; P_{ii}^{(k)} > 0\}$ is greater than 1), returns to a state can occur only at multiples of the period $\gcd\{N; P_{ii}^{(N)} > 0\}$.

Theorem A20 *If a Markov chain is irreducible and aperiodic, then there exists a* unique *limiting distribution* $\underline{\pi}$. *In this case, as* $n \to \infty$, $P^n \to \underline{\pi}'1$, *the matrix whose rows are all identically* $\underline{\pi}'$.

Generating Functions

Definition: Generating Function Let $a_0, a_1, a_2, \ldots$ be a finite or infinite sequence of real numbers. Suppose the power series

$$\mathcal{A}(t) = \sum_{i=0}^{\infty} a_i t^i$$

converges for all $-\epsilon < t < \epsilon$ for some value of $\epsilon > 0$. Then we say that the sequence has a *generating function* $\mathcal{A}(t)$.

Note: Every bounded sequence has a generating function since the series $\sum_{i=0}^{\infty} t^i$ converges whenever $|t| < 1$. Thus, discrete probability functions have generating functions. The generating function of a random variable X or its associated probability function $f_X(x) = P[X = x]$ is given by

$$\mathcal{F}_X(t) = \sum_x f_X(x) t^x = E(t^X)$$

Note that if the random variable has finite expected value, then this converges on the interval $t \in [-1, 1]$.

The advantage of generating functions is that they provide a transform of the original distribution to a space where many operations are made much easier. We will give examples of this later. The most important single property is that they are in one-to-one correspondence with distributions such that the series converges; for each distribution there is a unique generating function, and for each generating function, there is a unique distribution.

As a consequence of this representation and the following theorem, we can use generating functions to determine distributions that would otherwise be difficult to identify.

Theorem A21 *Suppose a random variable* X *has generating function* $\mathcal{F}_X(t)$ *and* Y *has generating function* $\mathcal{F}_Y(t)$. *Suppose that* X *and* Y *are independent.*

Then the generating function of the random variable $W = X + Y$ *is* $\mathcal{F}_W(t) = \mathcal{F}_X(t)\mathcal{F}_Y(t)$.

Notice that whenever a moment-generating function exists, we can recover the generating function from it by replacing e^t by t.

Example One of six different varieties of coupons is placed in each box of cereal. Find the distribution of the number of cereal boxes you need to buy to obtain all six coupons. Suppose X_1 is the number of boxes you need before you collect your first coupon, Then since $X_1 = 1$, the generating function X_1 is

$$E(t^{X_1}) = t.$$

Similarly if X_2 is the number of additional boxes required to obtain a new coupon, since $P(X_2 = j) = \left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^{j-1}$ the generating function of X_2 is

$$\sum_{j=1}^{\infty}\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^{j-1} t^j = \frac{5t}{6-t}$$

To obtain the third new coupon we will need X_3 boxes and X_3 has generating function

$$\sum_{j=1}^{\infty}\left(\frac{4}{6}\right)\left(\frac{2}{6}\right)^{j-1} t^j = \frac{4t}{6-2t}$$

Similarly we obtain the generating function of X_4, X_5, X_6 and since the total number of boxes required is the sum of the six independent random variables $X_1 + X_2 + ... + X_6$, it has generating function obtained as the product

$$t \times \frac{5t}{6-t} \times \frac{4t}{6-2t} \times ... \times \frac{t}{6-5t} = \frac{5!t^6}{(6-t)(6-2t)(6-3t)(6-4t)(6-5t)}$$
$$= \frac{5}{324}t^6 + \frac{25}{648}t^7 + \frac{175}{2916}t^8 + \frac{875}{11\,664}t^9 + \frac{11\,585}{139\,968}t^{10} + \frac{875}{10\,368}t^{11} + O\left(t^{12}\right)$$

from which we discover that the probability of only six cereal boxes is $\frac{5}{324}$, the probability of seven is $\frac{25}{648}$, and so on.

The Poisson Process

One of the simplest continuous-time stochastic processes is the *Poisson process*. Suppose N_t denotes the total number of arrivals into a system (such as the number of customers arriving at a queue) until time t. Note that the number of arrivals in time interval $(a, b]$ is then $N_b - N_a$. Assume the following properties:

(a) The probability of exactly one arrival in a small interval of length Δt is $\lambda \Delta t + o(\Delta t)$. (Note that the probability does not depend on where the interval is, only on its length.)

(b) The probability of two or more arrivals in an interval of length Δt is $o(\Delta t)$, where by definition of the o notation, $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$.

(c) For disjoint intervals $I_i = (a_i, b_i]$ (so $I_i \cap I_j = \phi, i \neq j$), the numbers of arrivals in these intervals $Y_i = N_{b_i} - N_{a_i}$ are mutually independent random variables.

Theorem A22 *Under the above conditions, (a)–(c), the distribution of the process* $N_t, t \in T$ *is that of a* Poisson process. *This means that the number of arrivals* $N_b - N_a$ *in an interval* $(a, b]$ *has a Poisson distribution with parameter* $\lambda(b-a) = \lambda \times$ *the* length of the interval, *and the number of arrivals in disjoint time intervals are independent random variables. The parameter* λ *specifies the* rate *of the Poisson process.*

We can easily show that if $N(t)$ is a Poisson process and $T_1, T_2, \ldots$ are the time of the first event, the time between the first and second events, and so on, then $T_1, T_2, \ldots$ are independent random variables, each with an exponential distribution with expected value $1/\lambda$. Moreover, if $T_1, T_2, \ldots, T_n$ are independent random variables each with an exponential(1) distribution, then the sum $\sum_{i=1}^{n} T_i$ has a (gamma) probability density function with probability density function

$$f(x) = \frac{1}{(n-1)!} x^{n-1} e^{-x}, \qquad \text{for } x > 0$$

This means that the event times for a Poisson process are gamma distributed.

Poisson Process in Space In an analogous way we may define a Poisson process in space as a distribution governing the occurrence of random points with the properties indicated above: The number of points in a given set S has a Poisson distribution with parameter $\lambda \times |S|$, where $|S|$ is the area or volume of the set, and if $Y_1, Y_2, \ldots$ are the number of points occurring in disjoint sets $S_1, S_2, \ldots$, they are mutually independent random variables.

2

Conditional Expectation and Martingales

2.1 CONDITIONAL EXPECTATION FOR SQUARE INTEGRABLE RANDOM VARIABLES

Information in probability and its applications is related to the notion of sigma algebras. For example, if I wish to predict whether tomorrow will be wet or dry ($X_2 = 1$ or 0) based only on similar results for today (X_1) and yesterday (X_0), then I am restricted to random variables that are functions $g(X_0, X_1)$ of the state on these two days. In other words, the random variable must be measurable with respect to the sigma algebra generated by X_0, X_1. Our objective is, in some sense, to get as close as possible to the unobserved value of X_2 using only random variables that are measurable with respect to this sigma algebra. This is essentially one way of defining conditional expectation. It provides the closest approximation to a random variable X if we restrict ourselves to random variables Y measurable with respect so some coarser sigma algebra.

Conditional Expectation

Theorem A23 *Let $\mathcal{G} \subset \mathcal{F}$ be sigma algebras and X a random variable on $(\Omega, \mathcal{F}, P)$. Assume $E(X^2) < \infty$. Then there exists an almost surely unique $\mathcal{G}$-measurable Y such that*

$$E[(X - Y)^2] = \inf_Z E(X - Z)^2 \tag{2.1}$$

where the infimum (greatest lower bound) is over all $\mathcal{G}$-measurable random variables. Note: We denote the minimizing Y by $E(X|\mathcal{G})$.

For two such minimizing Y_1, Y_2 (i.e., random variables Y that satisfy (2.1)), we have $P[Y_1 = Y_2] = 1$. This implies that conditional expectation is almost surely unique.

Suppose $\mathcal{G} = \{\varphi, \Omega\}$. What is $E(X|\mathcal{G})$? What random variables are measurable with respect to $\mathcal{G}$? Any nontrivial random variable that takes two or more possible values generates a nontrivial sigma algebra that includes sets other than the empty set that are strict subsets of the probability space Ω. Only a constant random variable is measurable with respect to the trivial sigma algebra $\mathcal{G}$. So the question becomes, what constant is as close as possible to all of the values of the random variable X in the sense of mean squared error? The obvious answer is the correct one, the expected value of X, because this leads to the same minimization discussed before, $\min_c E[(X-c)^2] = \min_c\{\text{var}(X) + (EX-c)^2\}$, which results in $c = E(X)$.

Example Suppose $\mathcal{G} = \{\varphi, A, A^c, \omega\}$ for some event A. What is $E(X|\mathcal{G})$?

Consider a candidate random variable Z taking the value a on A and b on the set A^c. Then

$$\begin{aligned} E[(X-Z)^2] &= E[(X-a)^2 I_A] + E[(X-b)^2 I_{A^c}] \\ &= E(X^2 I_A) - 2aE(XI_A) + a^2 P(A) \\ &\quad + E(X^2 I_{A^c}) - 2bE(XI_{A^c}) + b^2 P(A^c) \end{aligned}$$

Minimizing this with respect to both a and b results in

$$\begin{aligned} a &= E(XI_A)/P(A) \\ b &= E(XI_{A^c})/P(A^c) \end{aligned}$$

These values a and b are usually referred to in elementary probability as $E(X|A)$ and $E(X|A^c)$, respectively. Thus, the conditional expected value can be written

$$E(X|\mathcal{G})(\omega) = \begin{cases} E(X|A) & \text{if } \omega \in A \\ E(X|A^c) & \text{if } \omega \in A^c \end{cases}$$

As a special case consider X to be an indicator random variable $X = I_B$. Then we usually denote $E(I_B|\mathcal{G})$ by $P(B|\mathcal{G})$ and

$$P(B|\mathcal{G})(\omega) = \begin{cases} P(B|A) & \text{if } \omega \in A \\ P(B|A^c) & \text{if } \omega \in A^c \end{cases}$$

Note: Expected value is a constant, but the conditional expected value $E(X|\mathcal{G})$is a random variable measurable with respect to $\mathcal{G}$. Its value on the atoms (the distinct elementary subsets) of $\mathcal{G}$ is the average of the random variable X over these atoms.

Example Suppose $\mathcal{G}$ is generated by a finite partition $\{A_1, A_2, \ldots, A_n\}$ of the probability space Ω. What is $E(X|\mathcal{G})$?

In this case, any $\mathcal{G}$-measurable random variable is constant on the sets in the partition $A_j, j = 1, 2, \ldots, n$, and an argument similar to the one above shows that the conditional expectation is the simple random variable

$$E(X|\mathcal{G})(\omega) = \sum_{i=1}^{n} c_i I_{A_i}(\omega) \quad \text{where } c_i = E(X|A_i) = \frac{E(XI_{A_i})}{P(A_i)}$$

Example Consider the probability space $\Omega = (0, 1]$ together with $P =$ Lebesgue measure and the Borel sigma algebra. Suppose the function $X(\omega)$ is Borel measurable. Assume that $\mathcal{G}$ is generated by the intervals $(\frac{j-1}{n}, \frac{j}{n}]$ for $j = 1, 2, \ldots, n$. What is $E(X|\mathcal{G})$?

In this case

$$E(X|\mathcal{G})(\omega) = n \int_{(j-1)/n}^{j/n} X(s)ds \quad \text{when } \omega \in \left(\frac{j-1}{n}, \frac{j}{n}\right]$$
$$= \text{average of } X \text{ values over the relevant interval}$$

Theorem A24 (Properties of Conditional Expectation)

(a) *If a random variable X is $\mathcal{G}$-measurable, $E(X|\mathcal{G}) = X$.*

(b) *If a random variable X is independent of a sigma algebra $\mathcal{G}$, then $E(X|\mathcal{G}) = E(X)$.*

(c) *For any square integrable $\mathcal{G}$-measurable Z, $E(ZX) = E[ZE(X|\mathcal{G})]$.*

(d) (special case of (c)) $\int_A X\, dP = \int_A E(X|\mathcal{G}]dP$ *for all* $A \in \mathcal{G}$.

(e) $E(X) = E[E(X|\mathcal{G})]$.

(f) *If a $\mathcal{G}$-measurable random variable Z satisfies $E[(X - Z)Y] = 0$ for all other $\mathcal{G}$-measurable random variables Y, then $Z = E(X|\mathcal{G})$.*

(g) *If Y_1, Y_2 are distinct $\mathcal{G}$-measurable random variables both minimizing $E(X - Y)^2$, then $P(Y_1 = Y_2) = 1$.*

(h) *Additive: $E(X + Y|\mathcal{G}) = E(X|\mathcal{G}) + E(Y|\mathcal{G})$.*
Linearity: $E(cX + d|\mathcal{G}) = cE(X|\mathcal{G}) + d$.

(i) *If Z is $\mathcal{G}$-measurable, $E(ZX|\mathcal{G}) = ZE(X|\mathcal{G})$ a.s.*

(j) *If $\mathcal{H} \subset \mathcal{G}$ are sigma algebras, $E[E(X|\mathcal{G})|\mathcal{H}] = E(X|\mathcal{H})$.*

(k) *If $X \le Y$, $E(X|\mathcal{G}) \le E(Y|\mathcal{G})$ a.s.*

(l) *Conditional Lebesgue dominated convergence. If $X_n \to X$ a.s. and $|X_n| \le Y$ for some integrable random variable Y, then $E(X_n|\mathcal{G}) \to E(X|\mathcal{G})$ in distribution.*

Note: In general, we define $E(X|Z) = E(X|\sigma(Z))$, the conditional expected value given the sigma algebra generated by X, $\sigma(X)$. We can define the conditional variance $\mathrm{var}(X|\mathcal{G}) = E\{(X - E(X|\mathcal{G}))^2|\mathcal{G}\}$.

Proof.

(a) Notice that for any random variable Z that is $\mathcal{G}$-measurable, $E(X-Z)^2 \geq E(X-X)^2 = 0$, and so the minimizing Z is X (by definition, this is $E(X|\mathcal{G})$).

(b) Consider a random variable Y measurable with respect $\mathcal{G}$ and therefore independent of X. Then

$$\begin{aligned} E(X-Y)^2 &= E[(X-EX+EX-Y)^2] \\ &= E[(X-EX)^2] + 2E[(X-EX)(EX-Y)] + E[(EX-Y)^2] \\ &= E[(X-EX)^2] + E[(EX-Y)^2] \quad \text{by independence} \\ &\geq E[(X-EX)^2] \end{aligned}$$

It follows that $E(X-Y)^2$ is minimized when we choose $Y = EX$, and so $E(X|\mathcal{G}) = E(X)$.

(c) For any $\mathcal{G}$-measurable square integrable random variable Z, we may define a quadratic function of λ by

$$g(\lambda) = E[(X - E(X|\mathcal{G}) - \lambda Z)^2]$$

By definition of $E(X|\mathcal{G})$, this function is minimized over all real values of λ at the point $\lambda = 0$, and therefore $g'(0) = 0$. Setting its derivative $g'(0) = 0$ results in the equation

$$E(Z(X - E(X|\mathcal{G}))) = 0$$

or $E(ZX) = E[ZE(X|\mathcal{G})]$.

(d) If in (c) we put $Z = I_A$, where $A \in \mathcal{G}$, we obtain $\int_A X\,dP = \int_A E(X|\mathcal{G}]dP$.

(e) Again, this is a special case of property (c) corresponding to $Z = 1$.

(f) Suppose a $\mathcal{G}$-measurable random variable Z satisfies $E[(X-Z)Y] = 0$ for all other $\mathcal{G}$-measurable random variables Y. Consider in particular $Y = E(X|\mathcal{G}) - Z$ and define

$$\begin{aligned} g(\lambda) &= E[(X - Z - \lambda Y)^2] \\ &= E((X-Z)^2 - 2\lambda E[(X-Z)Y] + \lambda^2 E(Y^2) \\ &= E(X-Z)^2 + \lambda^2 E(Y^2) \\ &\geq E(X-Z)^2 = g(0) \end{aligned}$$

In particular, $g(1) = E[(X - E(X|\mathcal{G}))^2] \geq g(0) = E(X-Z)^2$, and by the uniqueness of conditional expectation in Theorem A23, $Z = E(X|\mathcal{G})$ almost surely.

(g) This is just deja vu (Theorem A23) all over again.

(h) Consider, for an arbitrary $\mathcal{G}$-measurable random variable Z,

$$E[Z(X+Y-E(X|\mathcal{G})-E(Y|\mathcal{G}))] = E[Z(X-E(X|\mathcal{G}))] + E[Z(Y-E(Y|\mathcal{G}))] = 0 \quad \text{by property (c).}$$

It therefore follows from property (f) that $E(X+Y|\mathcal{G}) = E(X|\mathcal{G}) + E(Y|\mathcal{G})$.

By a similar argument we may prove $E(cX+d|\mathcal{G}) = cE(X|\mathcal{G}) + d$.

(i)–(l) We leave the proof of these properties as exercises ■

2.2 CONDITIONAL EXPECTATION FOR INTEGRABLE RANDOM VARIABLES

We have defined conditional expectation as a projection (i.e., a $\mathcal{G}$-measurable random variable that is the closest to X) only for random variables with finite variance. It is fairly easy to extend this definition to random variables X on a probability space $(\Omega, \mathcal{F}, P)$ that are integrable (i.e., for which $E(|X|) < \infty$). We wish to define $E(X|\mathcal{G})$ where the sigma algebra $\mathcal{G} \subset \mathcal{F}$. First, for nonnegative integrable X, we may choose a sequence of simple random variables $X_n \uparrow X$. Since simple random variables have only finitely many values, they have finite variance, and we can use the definition above for their conditional expectation. Then $E(X_n|\mathcal{G})$ is an increasing sequence of random variables and so it converges. Define $E(X|\mathcal{G})$ to be the limit. In general, for random variables taking positive and negative values, we define $E(X|\mathcal{G}) = E(X^+|\mathcal{G}) - E(X^-|\mathcal{G})$. There are a number of details that need to be ironed out. First we need to show that this new definition is consistent with the old one when the random variable happens to be square integrable. We can also show that properties (a)–(i) above all hold under this new definition of conditional expectation. We close with the more common definition of conditional expectation found in most probability and measure theory texts, essentially property (d) above. It is, of course, equivalent to the definition as a projection that we used above when the random variable is square integrable, and when it is only integrable, it reduces to the aforementioned limit of the conditional expectations of simple functions.

Theorem A25 Consider a random variable X defined on a probability space $(\Omega, \mathcal{F}, P)$ for which $E(|X|) < \infty$. Suppose the sigma algebra $\mathcal{G} \subset \mathcal{F}$. Then there is a unique (almost surely P) $\mathcal{G}$-measurable random variable Z satisfying

$$\int_A X\,dP = \int_A Z\,dP \quad \text{for all } A \in \mathcal{G}$$

Any such Z we call the conditional expectation and denote by $E(X|\mathcal{G})$.

2.3 MARTINGALES IN DISCRETE TIME

In this section all random variables are defined on the same probability space $(\Omega, \mathcal{F}, P)$. Partial information about these random variables may be obtained from the observations so far, and in general, the "history" of a process up to time t is expressed through a sigma algebra $H_t \subset \mathcal{F}$. We are interested in stochastic processes or sequences of random variables called martingales—intuitively, the total fortune of an individual participating in a "fair game." In order for the game to be "fair," the expected value of your future fortune given the history of the process up to and including the present should be equal to your present wealth. In a sense you are tending to neither increase nor decrease your wealth over time; any fluctuations are purely random. Suppose your fortune at time s is denoted X_s. The values of the process of interest and any other related processes up to time s generate a sigma algebra H_s. Then the assertion that the game is fair implies that the expected value of our future fortune given this history of the process up to the present is exactly our present wealth $E(X_t|H_s) = X_s$ for $t > s$. In what follows, we will sometimes state our definitions to cover the discrete-time case in which t ranges through the integers $\{0, 1, 2, 3, \ldots\}$ or a subinterval of the real numbers such as $T = [0, \infty)$. In either case, T represents the set of possible indices t.

Definition $\{(X_t, H_t); t \in T\}$ is a *martingale* if

(a) H_t is an increasing (in t) family of sigma algebras.
(b) Each X_t is H_t-measurable and $E|X_t| < \infty$.
(c) For each $s < t$, where $s, t \in T$, we have $E(X_t|H_s) = X_s$ *a.s.*

Example Suppose Z_t are independent random variables with expectation 0. Define $H_t = \sigma(Z_1, Z_2, \ldots, Z_t)$ for $t = 1, 2, \ldots$ and $S_t = \sum_{i=1}^{t} Z_i$. Then notice that for integer $s < t$,

$$\begin{aligned} E[S_t|H_s] &= E[\sum_{i=1}^{t} Z_i|H_s] \\ &= \sum_{i=1}^{t} E[Z_i|H_s] \\ &= \sum_{i=1}^{s} Z_i \end{aligned}$$

because $E[Z_i|H_s] = Z_i$ if $i \le s$ and otherwise, if $i > s$, $E[Z_i|H_s] = 0$. Therefore, $\{(S_t, H_t),\ t = 1, 2, \ldots, \}$ is a (discrete-time) martingale. As an exercise you might show that if $E(Z_t^2) = \sigma^2 < \infty$, then $\{(S_t^2 - t\sigma^2, H_t), t = 1, 2, \ldots\}$ is also a discrete-time martingale.

Example Let X be any integrable random variable, and H_t an increasing family of sigma algebras for t in some index set T. Set $X_t = E(X|H_t)$. Then notice that for $s < t$,

$$E[X_t|H_s] = E[E[X|H_t]|H_s] = E[X|H_s] = X_s$$

So (X_t, H_t) is a martingale.

Technically, a sequence or set of random variables is not a martingale unless each random variable X_t is integrable. Of course, unless X_t is integrable, the concept of conditional expectation $E[X_t|H_s]$ is not even defined. You might think of reasons in each of the above two examples why the random variables X_t above and S_t in the previous example are indeed integrable.

Definition Let $\{(M_t, H_t); t = 1, 2, \ldots\}$ be a martingale and A_t a sequence of random variables measurable with respect to $H_{t-1.}$ Then the sequence A_t is called **non-anticipating** (an alternative term is **predictable**, but this will have a slightly different meaning in continuous time).

In gambling, we must determine our stakes and our strategy on the tth play of a game based on the information available to use at time $t - 1$. Similarly, in investment, we must determine the weights on various components in our portfolio at the end of day (or hour or minute) $t - 1$ *before* the random marketplace determines our profit or loss for that period of time. In this sense, both gambling and investment strategies must be determined by non-anticipating sequences of random variables (although both gamblers and investors often dream otherwise).

Definition: Martingale Transform Let $\{(M_t, H_t), t = 0, 1, 2, \ldots\}$ be a martingale, and let A_t be a bounded non-anticipating sequence with respect to H_t. Then the sequence

$$\tilde{M}_t = A_1(M_1 - M_0) + \cdots + A_t(M_t - M_{t-1}) \tag{2.2}$$

is called a *martingale transform* of M_t.

The martingale transform is sometimes denoted $A \circ M$, and it is one simple transformation that preserves the martingale property.

Theorem A26 *The martingale transform $\{(\tilde{M}_t, H_t), t = 1, 2, \ldots\}$ is a martingale.*

Proof.

$$\begin{aligned} E[\widetilde{M}_j - \widetilde{M}_{j-1}|H_{j-1}] &= E[A_j(M_j - M_{j-1}|H_{j-1}] \\ &= A_j E[(M_j - M_{j-1}|H_{j-1}] \quad \text{since } A_j \text{ is } H_{j-1}\text{-measurable} \\ &= 0 \; a.s. \end{aligned}$$

Therefore,

$$E[\widetilde{M}_j|H_{j-1}] = \widetilde{M}_{j-1} a.s. \qquad \blacksquare$$

Consider a random variable τ that determines when we stop betting or investing. Its value can depend arbitrarily on the outcomes in the past, as long as the decision to stop at time $\tau = t$ depends only on the results at time $t, t-1, \ldots$. Such a random variable is called an optional stopping time.

Definition A random variable τ taking values in $\{0, 1, 2, \ldots\} \cup \{\infty\}$ is a (optional) *stopping time* for a martingale $\{(X_t, H_t), t = 0, 1, 2, \ldots\}$ if for each n, $[\tau \leq t] \in H_t$.

If we stop a martingale at some random stopping time, the result continues to be a martingale, as the following theorem shows.

Theorem A27 *Suppose $\{(M_t, H_t), t = 1, 2, \ldots\}$ is a martingale and τ is an optional stopping time. Define a new sequence of random variables $Y_t = M_{t\wedge\tau} = M_{\min(t,\tau)}$ for $t = 0, 1, 2, \ldots$.Then $\{(Y_t, H_t), t = 1, 2, \ldots\}$ is a martingale.*

Proof. Notice that

$$M_{t\wedge\tau} = M_0 + \sum_{j=1}^{t}(M_j - M_{j-1})I(\tau \geq j)$$

Letting $A_j = I(\tau \geq j)$, this is a bounded H_{j-1}-measurable sequence and therefore $\sum_{j=1}^{n}(M_j - M_{j-1})I(\tau \geq j)$ is a martingale transform. By Theorem A26, it is a martingale. $\blacksquare$

Example (Ruin Probabilities) A random walk is a sequence of partial sums of the form $S_n = S_0 + \sum_{i=1}^{n} X_i$ where the random variables X_i are independent identically distributed. Suppose $P(X_i = 1) = p$, $P(X_i = -1) = q$, $P(X_i = 0) = 1 - p - q$ for $0 < p + q \leq 1$, and $p \neq q$. This is a model for our total fortune after we play n games, each game independent, and resulting either in a win of \$1, a loss of \$1, or break-even (no money changes hands). However we assume that the game is not fair, so that the probability of a win and the probability of a loss are different. We can show that

$$M_t = (q/p)^{S_t}, \quad t = 0, 1, 2, \ldots$$

is a martingale with respect to the usual history process $H_t = \sigma(X_1, Z_2, \ldots, X_t)$. Suppose that our initial fortune lies in some interval $A < S_0 < B$ and define the optional stopping time τ as the first time we hit either of two barriers at A or B. Then $M_{t \wedge \mathcal{T}}$ is a martingale. Suppose we wish to determine the probability of hitting the two barriers A and B in the long run. Since $E(M_{\mathcal{T}}) = \lim_{t\to\infty} E(M_{t\wedge\mathcal{T}}) = (q/p)^{S_0}$, by dominated convergence we have

$$(q/p)^A p_A + (q/p)^B p_B = (q/p)^{S_0} \tag{2.3}$$

where p_A and $p_B = 1 - p_A$ are the probabilities of hitting absorbing barriers at A or B, respectively. Solving, it follows that

$$((q/p)^A - (q/p)^B)p_A = (q/p)^{S_0} - (q/p)^B \tag{2.4}$$

or that

$$p_A = \frac{(q/p)^{S_0} - (q/p)^B}{(q/p)^A - (q/p)^B} \tag{2.5}$$

In the case $p = q$, a similar argument provides

$$p_A = \frac{B - S_0}{B - A} \tag{2.6}$$

These are often referred to as ruin probabilities and are of critical importance in the study of the survival of financial institutions such as insurance firms.

Definition For an optional stopping time τ, define the sigma algebra corresponding to the history up to the stopping time $H_{\mathcal{T}}$ to be the set of all events $A \in H$ for which

$$A \cap [\tau \le t] \in H_t \quad \text{for all } t \in T \tag{2.7}$$

Theorem A28 *$H_{\mathcal{T}}$ is a sigma-algebra.*

Proof. Clearly, since the empty set $\varphi \in H_t$ for all t, the same applies $\varphi \cap [\tau \le t]$ and so $\varphi \in H_{\mathcal{T}}$. We also need to show that if $A \in H_{\mathcal{T}}$, then the same applies the complement A^c. Notice that for each n,

$$\begin{aligned}[\tau \le t] \cap & \{A \cap [\tau \le t]\}^c \\ & = [\tau \le t] \cap \{A^c \cup [\tau > t]\} \\ & = A^c \cap [\tau \le t]\end{aligned}$$

and since each of the sets $[\tau \le t]$ and $A \cap [\tau \le t]$ are H_t-measurable, so must be the set $A^c \cap [\tau \le t]$. Since this holds for all t, it follows that whenever $A \in H_T$, then so it is for A^c. Finally, consider a sequence of sets $A_m \in H_T$ for all $m = 1, 2, \ldots$. We need to show that the countable union $\cup_{m=1}^{\infty} A_m \in H_T$. But

$$\{\cup_{m=1}^{\infty} A_m\} \cap [\tau \le t] = \cup_{m=1}^{\infty}\{A_m \cap [\tau \le t]\}$$

and by assumption the sets $\{A_m \cap [\tau \le t]\} \in H_t$ for each t. Therefore,

$$\cup_{m=1}^{\infty}\{A_m \cap [\tau \le t]\} \in H_t$$

and since this holds for all t, $\cup_{m=1}^{\infty} A_m \in H_T$. ■

There are several generalizations of the notion of a martingale that are quite common. In general, they modify the strict rule that the conditional expectation of the future given the present $E[X_t|H_s]$ is exactly equal to the present value X_s for $s < t$. One, a submartingale, models a process in which the conditional expectation satisfies an inequality compatible with a game that is either fair or is in your favor so that your fortune is expected either to remain the same or to increase.

Definition $\{(X_t, H_t); t \in T\}$ is a *submartingale* if

(a) H_t is an increasing (in t) family of sigma algebras.
(b) Each X_t is H_t-measurable and $E|X_t| < \infty$.
(c) For each $s < t$, $E(X_t|H_s) \ge X_s$ *a.s.*

Note that every martingale is a submartingale.

There is a very useful inequality, Jensen's inequality, referred to in most elementary probability texts. Consider a real-valued function $\phi(x)$ with the property that for any $0 < p < 1$, and for any two points x_1, x_2 in the domain of the function, the inequality

$$\phi(px_1 + (1-p)x_2) \le p\phi(x_1) + (1-p)\phi(x_2)$$

holds. Roughly, this says that the function evaluated at the average is less than the average of the function at the two endpoints or that the line segment joining the two points $(x_1, \phi(x_1))$ and $(x_2, \phi(x_2))$ lies above or on the graph of the function everywhere. Such a function is called a *convex function*. Functions like $\phi(x) = e^x$ and $\phi(x) = x^p, p \geq 1$, are convex functions, but $\phi(x) = \ln(x)$ and $\phi(x) = \sqrt{x}$ are not convex (in fact, they are concave). Notice that if a random variable X took two possible values x_1, x_2 with probabilities $p, 1 - p$, respectively, then this inequality asserts that the function at the point $E(X)$ is less than or equal $E\phi(X)$,

$$\phi(EX) \leq E\phi(X)$$

There is also a version of Jensen's inequality for conditional expectation that generalizes this result, and we will prove this more general version.

Theorem A29 (Jensen's Inequality) *Let ϕ be a convex function. Then for any random variable X and sigma field H,*

$$\phi(E(X|H)) \leq E(\phi(X)|H) \tag{2.8}$$

Proof. Consider the set $\mathcal{L}$ of linear functions $L(x) = a + bx$ that lie entirely below the graph of the function $\phi(x)$. It is easy to see that for a convex function

$$\phi(x) = \sup\{L(x); L \in \mathcal{L}\}$$

For any such line,

$$\begin{aligned} E(\phi(X)|H) &\geq E(L(X)|H) \\ &\geq L(E(X)|H)) \end{aligned}$$

If we take the supremum over all $L \in \mathcal{L}$, we obtain

$$E(\phi(X)|H) \geq \phi(E(X)|H)) \qquad \blacksquare$$

The standard version of Jensen's inequality follows on taking H above to be the trivial sigma field. Now from Jensen's inequality we can obtain a relationship among various commonly used norms for random variables. Define the norm $||X||_p = \{E(|X|^p)\}^{1/p}$ for all $p \geq 1$. The norm allows us to measure distances between two random variables; for example, a distance between X and Y can be expressed as

$$||X - Y||_p$$

It is well known that

$$||X||_p \leq ||X||_q \quad \text{whenever } 1 \leq p < q \tag{2.9}$$

This is easy to show since the function $\phi(x) = |x|^{q/p}$ is convex provided that $q \geq p$, and by Jensen's inequality,

$$E(|X|^q) = E(\phi(|X|^p) \geq \phi(E(|X|^p)) = |E(|X|^p)|^{q/p}$$

A similar result holds when we replace expectations with conditional expectations. Let X be any random variable and H be a sigma field. Then for $1 \leq p \leq q < \infty$,

$$\{E(|X|^p|H)\}^{1/p} \leq \{E(|X|^q|H)\}^{1/q} \tag{2.10}$$

Proof. Consider the function $\phi(x) = |x|^{q/p}$. This function is convex provided that $q \geq p$, and by the conditional form of Jensen's inequality,

$$E(|X|^q|H) = E(\phi(|X|^p)|H) \geq \phi(E(|X|^p|H)) = |E(|X|^p|H)|^{q/p} \text{ a.s.} \quad \blacksquare$$

In the special case that H is the trivial sigma field, this is the inequality

$$||X||_p \leq ||X||_q \tag{2.11}$$

Theorem A30 (Constructing Submartingales) *Let $\{(S_t, H_t), t = 1, 2, ...\}$ be a martingale. Then $(|S_t|^p, H_t)$ is a submartingale for any $p \geq 1$ provided that $E|S_t|^p < \infty$ for all t. Similarly, $((S_t - a)^+, H_t)$ is a submartingale for any constant a.*

Proof. Since the function $\phi(x) = |x|^p$ is convex for $p \geq 1$, it follows from the conditional form of Jensen's inequality that

$$E(|S_{t+1}|^p|H_t) = E(\phi(S_{t+1})|H_t) \geq \phi(E(S_{t+1}|H_t)) = \phi(S_t) = |S_t|^p \text{ a.s.} \quad \blacksquare$$

Various other operations on submartingales will produce another submartingale. For example, if X_n is a submartingale and ϕ is a convex nondecreasing function with $E\phi(X_n) < \infty$, then $\phi(X_n)$ is a submartingale.

Theorem A31 (Doob's Maximal Inequality) *Suppose (M_n, H_n) is a nonnegative submartingale. Then for $\lambda > 0$ and $p \geq 1$,*

$$P\left(\sup_{0 \leq m \leq n} M_m \geq \lambda\right) \leq \lambda^{-p} E(M_n^p)$$

Proof. We prove this in the case $p = 1$. The general case we leave as a problem. Define a stopping time

$$\tau = \min\{m;\ M_m \geq \lambda\}$$

so that $\tau \leq n$ if and only if the maximum has reached the value λ by time n, or

$$P\left[\sup_{0\leq m\leq n} M_m \geq \lambda\right] = P[\tau \leq n]$$

Now on the set $[\tau \leq n]$, the maximum $M_\tau \geq \lambda$, so

$$\lambda I(\tau \leq n) \leq M_\tau I(\tau \leq n) = \sum_{i=1}^{n} M_i I(\tau = i) \tag{2.12}$$

By the submartingale property, for any $i \leq n$ and $A \in Hi$,

$$E(M_i I_A) \leq E(M_n I_A)$$

Therefore, taking expectations on both sides of (2.12), and noting that for all $i \leq n$

$$E(M_i I(\tau = i)) \leq E(M_n I(\tau = i))$$

we obtain

$$\lambda P(\tau \leq n) \leq E(M_n I(\tau \leq n)) \leq E(M_n)$$ ■

Once again define the norm $||X||_p = \{E(|X|^p)\}^{1/p}$. Then the following inequality permits a measure of the norm of the maximum of a submartingale.

Theorem A32 (Doob's L^p Inequality) *Suppose (M_n, H_n) is a nonnegative submartingale and set $M_n^* = \sup_{0\leq m\leq n} M_n$. Then for $p > 1$, and all n*

$$||M_n^*||_p \leq \frac{p}{p-1}||M_n||_p$$

One of the main theoretical properties of martingales is that they converge under fairly general conditions. Conditions are clearly necessary. For example, consider a simple random walk $S_n = \sum_{i=1}^n Z_i$, where Z_i are independent identically distributed with $P(Z_i = 1) = P(Z_i = -1) = \frac{1}{2}$. Starting with an arbitrary value of S_0, say $S_0 = 0$, this is a martingale, but as $n \to \infty$ it does not converge almost surely or in probability.

On the other hand, consider a Markov chain with the property $P\,(X_{n+1} = j|X_n = i) = \frac{1}{2i+1}$ for $j = 0, 1, \ldots, 2i$. Notice that this is a martingale, and beginning with a positive value, say $X_0 = 10$, it is a nonnegative martingale. Does it converge almost surely? If so, the only possible limit is $X = 0$ because the nature of the process is such that $P\,[|X_{n+1} - X_n| \geq 1|X_n = i] \geq \frac{2}{3}$ unless $i = 0$. Convergence to $i \neq 0$ is impossible since in that case there is a high probability of jumps of magnitude at least 1. However, X_n does converge almost surely, a consequence of the martingale convergence theorem. Does it converge in L_1, that is, in the sense that $E[|X_n - X|] \to 0$ as $n \to \infty$? If it did, then $E(X_n) \to E(X) = 0$, and this contradicts the martingale property of the sequence, which implies $E(X_n) = E(X_0) = 10$. This is an example of a martingale that converges almost surely but not in L_1.

Lemma A4 *If $(X_t, H_t), t = 1, 2, ..., n$, is a (sub)martingale and if α, β are optional stopping times with values in $\{1, 2, ..., n\}$ such that $\alpha \leq \beta$, then*

$$E(X_\beta|H_\alpha) \geq X_\alpha$$

with equality if X_t is a martingale.

Proof. It is sufficient to show that

$$\int_A (X_\beta - X_\alpha) dP \geq 0$$

for all $A \in H_\alpha$. Note that if we define $Z_i = X_i - X_{i-1}$ to be the submartingale differences, the submartingale condition implies

$$E(Z_j\,|H_i) \geq 0 \; a.s. \quad \text{whenever } i < j$$

Therefore, for each $j = 1, 2, \ldots, n$ and $A \in H_\alpha$,

$$\begin{aligned}\int_{A\cap[\alpha=j]} (X_\beta - X_\alpha) dP &= \int_{A\cap[\alpha=j]} \sum_{i=1}^{n} Z_i I(\alpha < i \leq \beta) dP \\ &= \int_{A\cap[\alpha=j]} \sum_{i=j+1}^{n} Z_i I(\alpha < i \leq \beta) dP \\ &= \int_{A\cap[\alpha=j]} \sum_{i=j+1}^{n} E(Z_i|H_{i-1}) I(\alpha < i) I(i \leq \beta) dP \\ &\geq 0 \quad a.s.\end{aligned}$$

since $I(\alpha < i), I(i \le \beta)$, and $A \cap [\alpha = j]$ are all measurable with respect to H_{i-1} and $E(Z_i|H_{i-1}) \ge 0$ *a.s.* If we add over all $j = 1, 2, \dots, n$, we obtain the desired result. ■

The following inequality is needed to prove a version of the submartingale convergence theorem.

Theorem A33 (Doob's Up-Crossing Inequality) *Let M_n be a submartingale and for $a < b$, define $N_n(a, b)$ to be the number of complete up-crossings of the interval (a, b) in the sequence $M_j, j = 0, 1, 2, \dots, n$. This is the largest k such that there are integers $i_1 < j_1 < i_2 < j_2 \dots < j_k \le n$ for which*

$$M_{i_l} \le a \quad \textit{and} \quad M_{j_l} \ge b \quad \textit{for all } l = 1, \dots, k$$

Then

$$(b - a)EN_n(a, b) \le E\{(M_n - a)^+ - (M_0 - a)^+\}$$

Proof. By Theorem A29, we may replace M_n by $X_n = (M_n - a)^+$ and this is still a submartingale. Then we wish to count the number of up-crossings of the interval $[0, b']$ where $b' = b - a$. Define stopping times for this process by $\alpha_0 = 0$.

$$\begin{aligned} \alpha_1 &= \min\{j; 0 \le j \le n, X_j = 0\} \\ \alpha_2 &= \min\{j; \alpha_1 \le j \le n, X_j \ge b'\} \\ &\vdots \\ \alpha_{2k-1} &= \min\{j; \alpha_{2k-2} \le j \le n, X_j = 0\} \\ \alpha_{2k} &= \min\{j; \alpha_{2k-1} \le j \le n, X_j \ge b'\} \end{aligned}$$

In any case, if α_k is undefined because we do not again cross the given boundary, we define $\alpha_k = n$. Now each of these random variables is an optional stopping time. If there is an up-crossing between X_{α_j} and $X_{\alpha_{j+1}}$ (where j is odd), then the distance traveled is

$$X_{\alpha_{j+1}} - X_{\alpha_j} \ge b'$$

If X_{α_j} is well defined (i.e., it is equal to 0) and there is no further up-crossing, then $X_{\alpha_{j+1}} = X_n$ and

$$X_{\alpha_{j+1}} - X_{\alpha_j} = X_n - 0 \ge 0$$

Similarly, if j is even, since by the above lemma $(X_{\alpha_j}, H_{\alpha_j})$ is a submartingale,

$$E(X_{\alpha_{j+1}} - X_{\alpha_j}) \ge 0$$

Adding over all values of j and using the fact that $\alpha_0 = 0$ and $\alpha_n = n$,

$$E\sum_{j=0}^{n}(X_{\alpha_{j+1}} - X_{\alpha_j}) \geq b'EN_n(a,b)$$

$$E(X_n - X_0) \geq b'EN_n(a,b)$$

In terms of the original submartingale, this gives

$$(b-a)EN_n(a,b) \leq E(M_n - a)^+ - E(M_0 - a)^+ \qquad \blacksquare$$

Doob's martingale convergence theorem that follows is one of the nicest results in probability and is one of the reasons martingales are so frequently used in finance, econometrics, clinical trials, and life testing.

Theorem A34 (Submartingale Convergence Theorem) *Let* (M_n, H_n), $n = 1, 2, \ldots$, *be a submartingale such that* $\sup_{n\to\infty} EM_n^+ < \infty$. *Then there is an integrable random variable* M *such that* $M_n \to M$ *a.s. If* $\sup_n E(|M_n|^p) < \infty$ *for some* $p > 1$, *then* $||M_n - M||_p \to 0$.

Proof. The proof is an application of the up-crossing inequality. Consider any interval $a < b$ with rational endpoints. By the up-crossing inequality,

$$E(N_a(a,b)) \leq \frac{1}{b-a}E(M_n - a)^+ \leq \frac{1}{b-a}[|a| + E(M_n^+)]. \qquad (2.13)$$

Let $N(a,b)$ be the total number of times that the martingale completes an up-crossing of the interval $[a,b]$ over the infinite time interval $[1,\infty)$, and note that $N_n(a,b) \uparrow N(a,b)$ as $n \to \infty$. Therefore, by monotone convergence $E(N_a(a,b)) \to EN(a,b)$ and by (2.13),

$$E(N(a,b)) \leq \frac{1}{b-a} \limsup [a + E(M_n^+)] < \infty$$

This implies

$$P[N(a,b) < \infty] = 1$$

Therefore,

$$P(\liminf M_n \leq a < b \leq \limsup M_n) = 0$$

for every rational $a < b$, and this implies that M_n converges almost surely to a (possibly infinite) random variable. Call this limit M. We need to show that this random variable is almost surely finite. Because $E(M_n)$ is nondecreasing,

$$E(M_n^+) - E(M_n^-) \geq E(M_0)$$

and so

$$E(M_n^-) \leq E(M_n^+) - E(M_0)$$

But by Fatou's lemma,

$$E(M^+) = E(\text{lim inf} M_n^+) \leq \text{lim inf} E M_n^+ < \infty$$

Therefore, $E(M^-) < \infty$, and consequently the random variable M is finite almost surely. The convergence in L^p norm follows from the results on uniform integrability of the sequence. ■

Theorem A35 (L^p Martingale Convergence Theorem) *Let* (M_n, H_n), $n = 1, 2, \ldots,$ *be a martingale such that* $\sup_{n\to\infty} E|M_n|^p < \infty, p > 1$. *Then there is a random variable M such that* $M_n \to M$ *a.s. and in* L^p.

Example (The Galton-Watson Process) Consider a population of Z_n individuals in generation n, each of which produces a random number ξ of offspring in the next generation so that the distribution of Z_{n+1} is that of $\xi_1 + \ldots + \xi_{Z_n}$ for independent identically distributed ξ. This process $Z_n, n = 1, 2, \ldots,$ is called the Galton-Watson process. Let $E(\xi) = \mu$. Assume we start with a single individual in the population $Z_0 = 1$ (otherwise, if there are j individuals in the population to start, then the population at time n is the sum of j independent terms, the offspring of each). Then the following properties hold:

1. The sequence Z_n/μ^n is a martingale.
2. If $\mu < 1$, $Z_n \to 0$ and $Z_n = 0$ for all sufficiently large n.
3. If $\mu = 1$ and $P(\xi \neq 1) > 0$, then $Z_n = 0$ for all sufficiently large n.
4. If $\mu > 1$, then $P(Z_n = 0 \text{ for some } n) = \rho$, where ρ is the unique value < 1 satisfying $E(\rho^\xi) = \rho$.

Definition: Supermartingale $\{(X_t, H_t); t \in T\}$ is a *supermartingale* if

(a) H_t is an increasing (in t) family of sigma algebras.
(b) Each X_t is H_t-measurable and $E|X_t| < \infty$.
(c) For each $s < t$, $s, t \in T$, $E(X_t|H_s) \leq X_s$ *a.s.*

The properties of supermartingales are very similar to those of submartingales, except that the expected value is a nonincreasing sequence. For example, if $A_n \geq 0$ is a predictable (non-anticipating) bounded sequence and (M_n, H_n) is a supermartingale, then the supermartingale transform $A \circ M$ is a supermartingale. Similarly, if in addition the supermartingale is nonnegative $M_n \geq 0$, then there is a random variable M such that $M_n \to M$ *a.s.* with $E(M) \leq E(M_0)$. The following example shows that a nonnegative supermartingale may converge almost surely and yet not converge in expected value.

Example Let S_n be a simple symmetric random walk with $S_0 = 1$ and define the optional stopping time $N = \{n; S_n = 0\}$. Then

$$X_n = S_{n \wedge N}$$

is a nonnegative (super)martingale, and therefore X_n converges almost surely. The limit (call it X) must be 0 because if $X_n > 0$ infinitely often, then $|X_{n+1} - X_n| = 1$ for infinitely many n, and this contradicts the convergence. However, in this case, $E(X_n) = 1$ whereas $E(X) = 0$, so the convergence is not in the L_1 norm (in other words, $||X - X_n||_1 \nrightarrow 0$) or in expected value.

A martingale under a reversal of the direction of time is a reverse martingale. The sequence $\{(X_t, H_t); t \in T\}$ is a *reverse martingale* if

(a) H_t is a decreasing (in t) family of sigma algebras.
(b) Each X_t is H_t-measurable and $E|X_t| < \infty$.
(c) For each $s < t$, $E(X_s|H_t) = X_t$ *a.s.*

It is easy to show that if X is any integrable random variable, and if H_t is any decreasing family of sigma algebras, then $X_t = E(X|H_t)$ is a reverse martingale. Reverse martingales require even fewer conditions than do martingales for almost sure convergence.

Theorem A36 (Reverse Martingale Convergence Theorem) *If (X_n, H_n), $n = 1, 2, \ldots,$ is a reverse martingale, then as $n \to \infty$, X_n converges almost surely to the random variable $E(X_1| \cap_{n=1}^{\infty} H_n)$.*

The reverse martingale convergence theorem can be used to give a particularly simple proof of the strong law of large numbers because if $Y_i, i = 1, 2, \ldots$, are independent identically distributed random variables and we define H_n to be the sigma algebra $\sigma(\bar{Y}_n, Y_{n+1}, Y_{n+2}, \ldots)$, where $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, then H_n is a decreasing family of sigma fields and $\bar{Y}_n = E(Y_1|H_n)$ is a reverse martingale.

2.4 UNIFORM INTEGRABILITY

Definition A set of random variables $\{X_i, i = 1, 2, \ldots\}$ is uniformly integrable if

$$\sup_i E(|X_i| I(|X_i| > c) \to 0 \quad \text{as } c \to \infty$$

Some Properties of Uniform Integrability

1. Any finite set of integrable random variables is uniformly integrable.
2. Any infinite sequence of random variables that converges in L^1 is uniformly integrable.
3. Conversely, if a sequence of random variables converges almost surely and is uniformly integrable, then it also converges in L^1.
4. If X is integrable on a probability space (Ω, H) and H_t is any family of sub-sigma fields, then $\{E(X|H_t)\}$ is uniformly integrable.
5. If $\{X_n, n = 1, 2, \ldots\}$ is uniformly integrable, then $\sup_n E(X_n) < \infty$.

Uniform integrability is the bridge between convergence almost surely or in probability, and convergence in expectation, as the following result shows.

Theorem A37 *Suppose a sequence of random variables satisfies $X_n \to X$ in probability. Then the following are all equivalent:*

1. $\{X_n, n = 1, 2, \ldots\}$ *is uniformly integrable.*
2. $X_n \to X$ *in* L^1.
3. $E(|X_n|) \to E(|X|)$.

As a result a uniformly integrable submartingale $\{X_n, n = 1, 2, \ldots\}$ not only converges almost surely to a limit X as $n \to \infty$, but it converges in expectation and in L^1 as well; in other words $E(X_n) \to E(X)$ and $E(|X_n - X|) \to 0$ as $n \to \infty$. There is one condition useful for demonstrating uniform integrability of a set of random variables:

Lemma A5 *Suppose there exists a function $\phi(x)$ such that $\lim_{x\to\infty} \phi(x)/x = \infty$ and $E\phi(|X_t|) \le B < \infty$ for all $t \ge 0$. Then the set of random variables $\{X_t; t \ge 0\}$ is uniformly integrable.*

One of the most common methods for showing uniform integrability, used in results such as the Lebesgue dominated convergence theorem, is to require that a sequence of random variables be dominated by a single integrable random variable X. This is, in fact, a special use of the above lemma because if X is an integrable random variable, then there exists a convex function $\phi(x)$ such that $\lim_{x\to\infty} \phi(x)/x = \infty$ and $E(\phi(|X|) < \infty$.

3

Stochastic Integration and Continuous-Time Models

3.1 BROWNIAN MOTION

The single most important continuous-time process in the construction of financial models is the Brownian motion process. Brownian motion is the oldest continuous-time model used in finance and goes back to Bachelier (1900), around the turn of the last century. It is also the most common building block for more sophisticated continuous-time models called diffusion processes.

The Brownian motion process is a random continuous-time process denoted $W(t)$ or W_t, defined for $t \geq 0$ such that $W(0)$ takes some predetermined value, usually 0, and for each $0 \leq s < t$, $W(t) - W(s)$ has a normal distribution with mean $\mu(t-s)$ and variance $\sigma^2(t-s)$. The parameters μ and σ are the drift and the diffusion parameters of the Brownian motion, and in the special case $\mu = 0, \sigma = 1$, $W(t)$ is often referred to as a standard Brownian motion or a Wiener process. Further properties of the Brownian motion process that are important are

- A Brownian motion process exists such that the sample paths are each continuous functions of t (with probability 1).
- The joint distribution of any finite number of increments $W(t_2) - W(t_1)$, $W(t_4) - W(t_3), \ldots, W(t_k) - W(t_{k-1})$are independent normal random variables provided that $0 \leq t_1 < t_2 \leq t_3 < t_4 \leq \cdots \leq t_{k-1} < t_k$.

Further Properties of the (Standard) Brownian Motion Process

The covariance between $W(t)$ and $W(s)$, $\text{cov}(W(t), W(s)) = \min(s, t)$. Brownian motion is an example of a *Gaussian process*, a process for which every finite-dimensional distribution such as $(W(t_1), W(t_2), ..., W(t_k))$ is normal (multivariate or univariate). In fact, Gaussian processes are uniquely determined by their covariance structure. In particular, if a Gaussian process has

$E(X_t) = 0$ and $\mathrm{cov}(X(t), X(s)) = \min(s, t)$, then it has independent increments. If in addition it has continuous sample paths and if $X_0 = 0$, then it is standard Brownian motion.

Toward the construction of a Brownian motion process, define the triangular function

$$\Delta(t) = \begin{cases} 2t & \text{for } 0 \le t \le \frac{1}{2} \\ 2(1-t) & \text{for } \frac{1}{2} \le t \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and similar functions with base of length 2^{-j}

$$\Delta_{j,k}(t) = \Delta(2^j t - k) \quad \text{for } j = 1, 2, \ldots, \quad k = 0, 1, \ldots, 2^j - 1$$
$$\Delta_{0,0}(t) = t, \quad 0 \le t \le 1$$

Theorem A38 (Wavelet Construction of Brownian Motion) *Suppose the random variables $Z_{j,k}$ are independent $N(0, 1)$ random variables. Then the series below converges uniformly (almost surely) to a standard Brownian motion process $B(t)$ on the interval* $[0, 1]$.

$$B(t) = \sum_{j=0}^{\infty} \sum_{k=0}^{2^j - 1} 2^{-j/2-1} Z_{j,k} \Delta_{j,k}(t)$$

The standard Brownian motion process can be extended to the whole interval $[0, \infty)$ by generating independent Brownian motion processes $B^{(n)}$ on

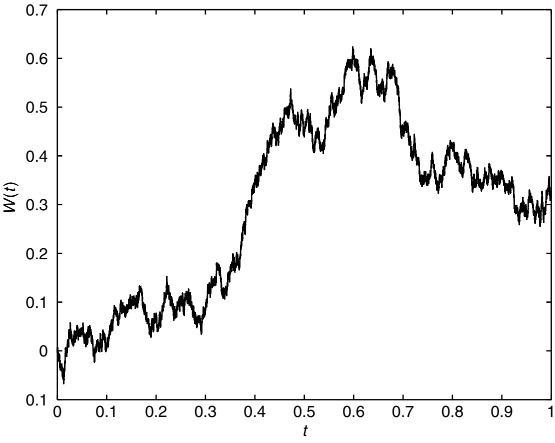

FIGURE 3.1 A Sample Path for the Standard Brownian Motion (Wiener) Process

the interval [0, 1] and defining $W(t) = \sum_{j=1}^{n} B^{(j)}(1) + B^{(n+1)}(t-n)$ whenever $n \le t < n+1$.

Figure 3.1 gives a sample path of the standard Brownian motion. Evidently the path is continuous, but if you examine it locally it appears to be just barely continuous, having no higher-order smoothness properties. For example, derivatives do not appear to exist because of the rapid fluctuations of the process everywhere. There are various modifications of the Brownian motion process that result in a process with exactly the same distribution.

Theorem A39 *If $W(t)$ is a standard Brownian motion process on $[0, \infty)$, then so are the processes $X_t = \frac{1}{\sqrt{a}} W(at)$ and $Y_t = tW(1/t)$ for any $a > 0$.*

A standard Brownian motion process is an example of a continuous-time martingale, because, for $s < t$,

$$\begin{aligned} E[W(t)|H_s] &= E[W(t) - W(s)|H_s] + E[W(s)|H_s] \\ &= 0 + W(s) \end{aligned}$$

since the increment $W(t) - W(s)$ is independent of the past and normally distributed with expected value 0. In fact, it is a continuous martingale in the sense that sample paths are continuous (with probability 1) functions of t. It is not the only continuous martingale, however. For example, it is not difficult to show that both $X_t = W_t^2 - t$ and $\exp(\alpha W_t - \alpha^2 t/2)$, for α any real number, are continuous martingales. Of course, neither is a Gaussian process. Their finite-dimensional distributions cannot be normal since both processes are restricted to values in the positive reals. We discussed earlier the ruin probabilities for a random walk using martingale theory, and a similar theory can be used to establish the boundary crossing probabilities for a Brownian motion process. The following theorem establishes the relative probability that a Brownian motion hits each of two boundaries, one above the initial value and the other below.

Theorem A40 (Ruin Probabilities for Brownian Motion) *If $W(t)$ is a standard Brownian motion and the stopping time τ is defined by*

$$\tau = \inf\{t;\ W(t) = -b \text{ or } a\}$$

where a and b are positive numbers, then $P(\tau < \infty) = 1$ and

$$P[W_T = a) = \frac{b}{a+b}$$

Although this establishes which boundary is hit with what probability, it says nothing about the time at which the boundary is first hit. The distribution of this hitting time (the first passage time distribution) is particularly simple:

Theorem A41 (Hitting Times for a Flat Boundary) *If $W(t)$ is a standard Brownian motion and the stopping time τ is defined by*

$$\tau_a = \inf\{t;\ W(t) = a\}$$

where $a > 0$, then

Theorem A42

1. $P(\tau_a < \infty) = 1$
2. τ_a *has a Laplace transform given by*

$$E(e^{-\lambda\tau_a}) = e^{-\sqrt{2\lambda}|a|}$$

3. *The probability density function of* τ_a *is*

$$f(t) = at^{-3/2}\phi(at^{-1/2})$$

where ϕ is the standard normal probability density function.

4. *The cumulative distribution function of* τ_a *is given by*

$$P[\tau_a \le t] = 2P[W(t) > a] = 2[1 - \Phi(at^{-1/2})] \text{ for } t > 0$$

and zero otherwise.

5. $E(\tau_a) = \infty$

The last property is surprising. The standard Brownian motion has no general tendency to rise or fall, but because of the fluctuations it is guaranteed to strike a barrier placed at any level $a > 0$. However, the time before this barrier is struck can be very long, so long that the expected time is infinite. The following corollary provides an interesting connection between the maximum of a Brownian motion over an interval and its value at the end of the interval.

Corollary If $W_t^* = \max\{W(s); 0 < s < t\}$, then for $a \ge 0$,

$$P[W_t^* > a] = P[\tau_a \le t] = 2P[W(t) > a]$$

Theorem A43 (Time of Last Return to 0) *Consider the random time* $L = \sup\{t \le 1; W(t) = 0\}$. *Then L has cumulative distribution function*

$$P[L \le s] = \frac{2}{\pi}\arcsin(\sqrt{s}), \quad 0 < s < 1$$

and corresponding probability density function

$$\frac{d}{ds}\frac{2}{\pi}\arcsin(\sqrt{s}) = \frac{1}{\pi\sqrt{s(1-s)}}, \quad 0 < s < 1$$

3.2 CONTINUOUS-TIME MARTINGALES

As usual, the value of the stochastic process at time t may be denoted by $X(t)$ or by X_t for $t \in [0, \infty)$, and let H_t be a sub-sigma field of H such that $H_s \subset H_t$ whenever $s \leq t$. We call such a sequence a *filtration*. X_t is said to be *adapted* to the filtration if $X(t)$ is H_t-measurable for all $t \in [0, \infty)$.

Henceforth, we assume that all stochastic processes under consideration are adapted to the filtration H_t. We also assume that the filtration H_t is *right continuous*, is, that

$$\bigcap_{\epsilon>0} H_{t+\epsilon} = H_t \tag{3.1}$$

We can make this assumption without loss of generality because if H_t is any filtration, then we can make it right continuous by replacing it with

$$H_{t+} = \bigcap_{\epsilon>0} H_{t+\epsilon} \tag{3.2}$$

We use the fact that the intersection of sigma fields is a sigma field. Note that any process that was adapted to the original filtration is also adapted to the new filtration H_{t+}. We also typically assume, by analogy to the definition of Lebesgue measurable sets, that if A is any set with $P(A) = 0$, then $A \in H_0$. These two conditions, that the filtration is right continuous and contains the $P-$null sets, are referred to as the *standard conditions*. The definition of a martingale is, in continuous time, essentially the same as in discrete time.

Definition Let $X(t)$ be a continuous-time stochastic process adapted to a right-continuous filtration H_t, where $0 \leq t < \infty$. Then X is a *martingale* if $E|X(t)| < \infty$ for all t and

$$E\left[X(t)|H_s\right] = X(s) \tag{3.3}$$

for all $s < t$. The process $X(t)$ is a *submartingale* (respectively, a *supermartingale*) if the equality is replaced by $\geq$ (respectively, $\leq$).

Definition A random variable τ taking values in $[0, \infty]$ is a *stopping time* for a martingale (X_t, H_t) if for each $t \geq 0$, the event $[\tau \leq t]$ is in the sigma algebra H_t.

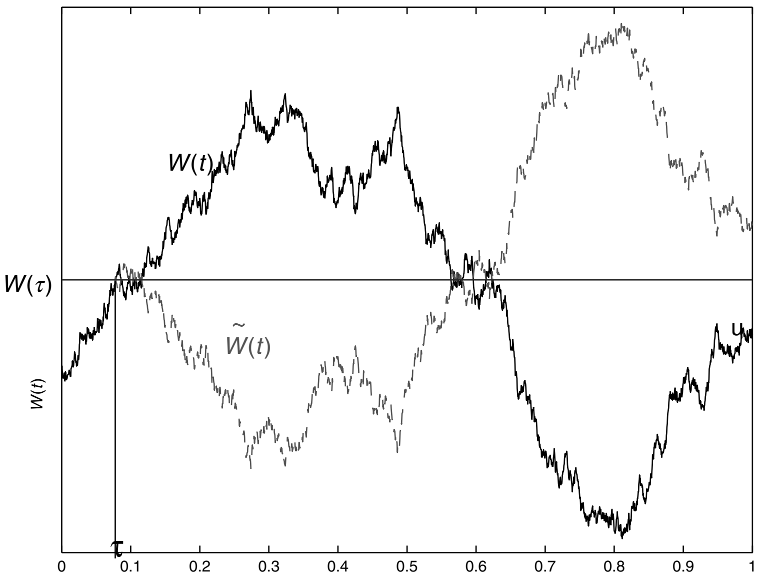

FIGURE 3.2 The Process $\widetilde{W}(t)$ Obtained by Reflecting a Brownian Motion about $W(\tau)$

Stopping a martingale at a sequence of nondecreasing stopping times preserves the martingale property, but there are some operations with Brownian motion that preserve the Brownian motion measure:

Theorem A44 (Reflection and Strong Markov Property) *If τ is a stopping time with respect to the usual filtration of a standard Brownian motion $W(t)$, then the process*

$$\widetilde{W}(t) = \begin{cases} W(t) & t < \tau \\ 2W(\tau) - W(t) & t \geq \tau \end{cases}$$

is a standard Brownian motion.

The process $\widetilde{W}(t)$ is obtained from the Brownian motion process as follows: Up to time τ the original Brownian motion is left alone, and for $t > \tau$, the process $\widetilde{W}(t)$ is the reflection of $W(t)$ about a horizontal line drawn at $y = W(\tau)$. This is shown in Figure 3.2.

Theorem A45 *Let $\{(M_t, H_t), t \geq 0\}$ be a (right-)continuous martingale and assume that the filtration satisfies the standard conditions. If τ is a stopping* *time, then the process*

$$X_t = M_{t\wedge T}$$

is also a continuous martingale with respect to the same filtration.

Various other results are essentially the same in discrete or continuous time. For example, Doob's L^p inequality

$$\left\|\sup_{0\le t\le T} M_t\right\|_p \le \frac{p}{p-1}||M_T||_p, \quad \text{if } p > 1$$

holds for right-continuous nonnegative submartingales and $p \geq 1$. Similarly, the submartingale convergence theorem holds as stated earlier, but with $n \to \infty$ replaced by $t \to \infty$.

3.3 INTRODUCTION TO STOCHASTIC INTEGRALS

The stochastic integral arose from attempts to use the techniques of Riemann-Stieltjes integration for stochastic processes. However, Riemann integration requires that the integrating function have *locally bounded variation* in order that the Riemann-Stieltjes sum converge.

Definition: Locally Bounded Variation If the process A_t can be written as the difference of two nondecreasing processes, it is called a *process of locally bounded variation*. A function is said to have locally bounded variation if it can be written as the difference of two nondecreasing processes.

For any function G of locally bounded variation, random or not, integrals such as $\int_0^T f\, dG$ are easy to define because, since we can write $G = G_1 - G_2$ as the difference between two nondecreasing functions G_1, G_2, the Rieman-Stieltjes sum

$$\sum_{i=1}^{n} f(s_i)[G(t_i) - G(t_{i-1})]$$

where $0 = t_0 < t_1 < t_2 < \cdots < t_n = T$ is a partition of $[0, T]$, and $t_{i-1} \leq s_i \leq t_i$ will converge to the same value regardless of where we place s_i in the interval (t_{i-1}, t_i) as the mesh size $\max_i |t_i - t_{i-1}| \to 0$.

By contrast, many stochastic processes do not have paths of bounded variation. Consider, for example, a hypothetical integral of the form

$$\int_0^T f\, dW$$

where f is a nonrandom function of $t \in [0, T]$ and W is a standard Brownian motion. The Riemann-Stieljes sum for this integral would be

$$\sum_{i=1}^{n} f(s_i)[W(t_i) - W(t_{i-1})]$$

where again $0 = t_0 < t_1 < t_2 < \cdots < t_n = T$, and $t_{i-1} \leq s_i \leq t_i$. In this case as $\max_i |t_i - t_{i-1}| \to 0$, the Riemann-Stieljes sum will not converge because the Brownian motion paths are not of bounded variation. When f has bounded variation, we can circumvent this difficulty by formally defining the integral using integration by parts. Thus if we formally write

$$\int_0^T f\, dW = f(T)W(T) - f(0)W(0) - \int_0^T W\, df$$

then the right-hand side is well defined and can be used as the definition of the left-hand side. Unfortunately, this simple interpretation of the stochastic integral does not work for many applications. The integrand f is often replaced by some function of W or another stochastic process that does not have bounded variation. There are other difficulties. For example, integration by parts to evaluate the integral

$$\int_0^T W\, dW$$

leads to $\int_0^T W\, dW = W^2(T)/2$, which is not the Ito stochastic integral. Suppose we return to the possible limiting values of the Riemann Stieltjes sums

$$I_\alpha = \sum_{i=1}^{n} f(s_i)\{W(t_i) - W(t_{i-1})\} \tag{3.4}$$

where $s_i = t_{i-1} + \alpha(t_i - t_{i-1})$ for some $0 \leq \alpha \leq 1$. If the Riemann integral were well defined, then $I_1 - I_0 \to 0$ in probability. However, when $f(s) = W(s)$,

$$I_1 - I_0 = \sum_{i=1}^{n} [W(t_i) - W(t_{i-1})]^2$$

and this cannot possibly converge to zero because, in fact, the expected value is

$$E\left(\sum_{i=1}^{n} [W(t_i) - W(t_{i-1})]^2\right) = \sum_{i=1}^{n} (t_i - t_{i-1}) = T$$

Since these increments $[W(t_i) - W(t_{i-1})]^2$ are independent, we can show by a version of the law of large numbers that

$$\sum_{i=1}^{n}[W(t_i) - W(t_{i-1})]^2 \rightarrow_p T$$

and more generally $I_\alpha - I_0 \rightarrow \alpha T$ in probability as the partition grows finer.

In other words, unlike the case for the Riemann-Stieltjes integral, it makes a difference where we place the point s_i in the interval (t_{i-1}, t_i) for a stochastic integral. The Ito stochastic integral corresponds to $\alpha = 0$ and approximates the integral $\int_0^T W\, dW$ with partial sums of the form

$$\sum_{i=1}^{n} W(t_{i-1})[W(t_i) - W(t_{i-1})]$$

the limit of which is, as the mesh size decreases, $\frac{1}{2}(W^2(T) - T)$. If we evaluate the integrand at the right endpoint of the interval (i.e., taking $\alpha = 1$), we obtain $\frac{1}{2}(W^2(T) + T)$. Another natural choice is $\alpha = 1/2$ (called the Stratonovich integral), and note that this definition gives the answer $W^2(T)/2$, which is the same result obtained from the usual Riemann integration by parts. Which definition is "correct"? The Stratonovich integral has the advantage that it satisfies most of the traditional rules of deterministic calculus; for example, *if the integral below is a Stratonovich integral,*

$$\int_0^T \exp(W_t) dW_t = \exp(W_T) - 1$$

While all definitions of a stochastic integral are useful, the main applications in finance are those in which the values $f(s_i)$ appearing in (3.4) are the weights on various investments in a portfolio, and the increment $[W(t_i) - W(t_{i-1})]$ represents the changes in price of the components of that portfolio over the next interval of time. Obviously, one must commit to one's investments *before* observing the changes in the values of those investments. For this reason the Ito integral ($\alpha = 0$) seems the most natural for these applications.

We now define the class of functions f to which this integral will apply. We assume that H_t is a standard Brownian filtration and that the interval $[0, T]$ is endowed with its Borel sigma field. Let $\mathcal{H}^2$ be the set of functions $f(\omega, t)$ on the product space $\Omega \times [0, T]$ such that

1. f is measurable with respect to the product sigma field on $\Omega \times [0, T]$.
2. For each $t \in [0, T]$, $f(., t)$ is H_t-measurable (in other words, the stochastic process $f(., t)$ is adapted to H_t).

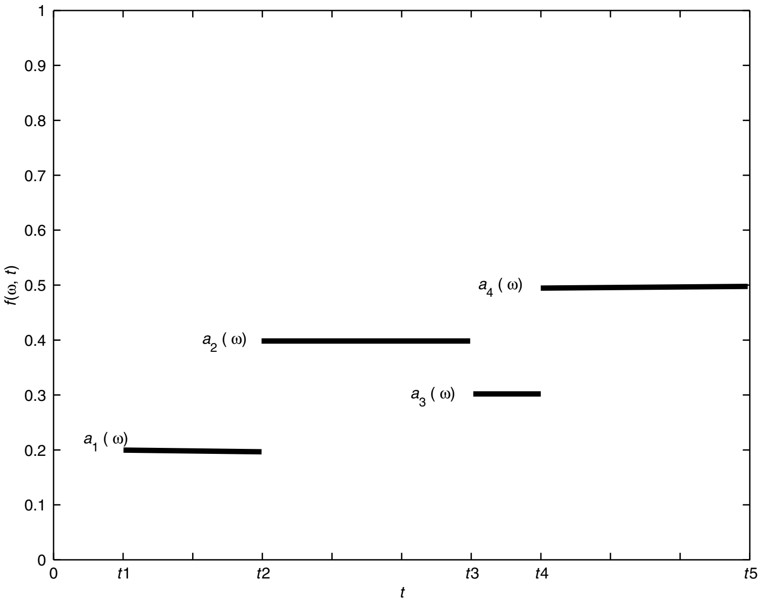

FIGURE 3.3 A Typical Step Function $f(\omega, t)$

3. $E[\int_0^T f^2(\omega, t)dt] < \infty$.

The set of processes $\mathcal{H}^2$ is the natural domain of the Ito integral. However, before we define the stochastic integral on $\mathcal{H}^2$, we need to define it in the obvious way on the step functions in $\mathcal{H}^2$. Let $\mathcal{H}_0^2$ be the subset of $\mathcal{H}^2$ consisting of functions of the form

$$f(\omega, t) = \sum_{i=0}^{n-1} a_i(\omega)\mathbf{1}(t_i < t \leq t_{i+1})$$

where the random variables a_i are measurable with respect to H_{t_i} and $0 = t_0 < t_1 < \cdots < t_n = T$. These functions f are predictable in that their value $a_i(\omega)$ in the interval $(t_i, t_{i+1}]$ is determined before we reach this interval. A typical step function is graphed in Figure 3.3.

For such functions, the stochastic integral has only one natural definition:

$$\int_0^T f(\omega, t)dW(t) = \sum_{i=0}^{n-1} a_i(\omega)(W(t_{i+1)} - W(t_i))$$

and note that considered as a function of T, this forms a continuous-time square integrable martingale.

There is a simple definition of an of inner product between two square integrable random variables X and Y, namely $E(XY)$, and we might ask how this inner product behaves when applied to the random variables obtained from stochastic integration, such as $X_T(\omega) = \int_0^T f(\omega, t)dW(t)$ and $Y_T(\omega) = \int_0^T g(\omega, t)dW(t)$. The answer is simple, in fact, and lies at the heart of Ito's definition of a stochastic integral. For reasons that will become a little clearer later, let us define the predictable covariation process to be the stochastic process described by

$$<X, Y>_T (\omega) = \int_0^T f(\omega, t)g(\omega, t)dt$$

Theorem A46 *For functions f and g in $\mathcal{H}_0^2$,*

$$E\{<X, Y>_T\} = E\{X_T Y_T\}. \tag{3.5}$$

and

$$E(<X, X>_T) = E\{\int_0^T f^2(\omega, t)dt\} = E(X_T^2) \tag{3.6}$$

These identities establish an isometry, a relationship between inner products, at least for two functions in $\mathcal{H}_0^2$. The norm on stochastic integrals defined by

$$\left\| \int_0^T f dW \right\|_{L(P)}^2 = E\left(\int_0^T f(\omega, t)dW(t) \right)^2$$

agrees with the usual L^2 norm on the space of random functions,

$$||f||_2^2 = E\left\{ \int_0^T f^2(\omega, t)dt \right\}$$

We use the notation $||f||_2^2 = E\{\int_0^T f^2(\omega, t)dt\}$. If we now wish to define a stochastic integral for a general function $f \in \mathcal{H}^2$, the method is fairly straightforward. First we approximate any $f \in \mathcal{H}^2$ using a sequence of step functions $f_n \in \mathcal{H}_0^2$ such that

$$||f - f_n||_2^2 \to 0$$

To construct the approximating sequence f_n, we can construct a mesh $t_i = \frac{i}{2^n}T$ for $i = 0, 1, \ldots, 2^n - 1$ and define

$$f_n(\omega, t) = \sum_{i=0}^{n-1} a_i(\omega)\mathbf{1}(t_i < t \le t_{i+1}) \tag{3.7}$$

with

$$a_i(\omega) = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} f(\omega, s)ds$$

the average of the function over the previous interval.

The definition of a stochastic integral for any $f \in \mathcal{H}^2$ is now clear from this approximation. Choose a sequence $f_n \in \mathcal{H}_0^2$ such that $||f - f_n||_2^2 \to 0$. Since the sequence f_n is Cauchy, the isometry property (3.6) shows that the stochastic integrals $\int_0^T f_n \, dW$ also forms a Cauchy sequence in $L_2(P)$. Since this space is complete (in the sense that Cauchy sequences converge to a random variable in the space), we can define $\int_0^T f \, dW$ to be the limit of the sequence $\int_0^T f_n \, dW$ as $n \to \infty$. Of course, there is some technical work to be done; for example, we need to show that two approximating sequences lead to the same integral and that the Ito isometry (3.5) still holds for functions f and g in $\mathcal{H}^2$. The details can be found in Steele (2001).

So far we have defined integrals $\int_0^T f \, dW$ for a fixed value of T, but how should we define the stochastic process $X_t = \int_0^t f \, dW$ for $t < T$? To do so we define a similar integral but with the function set to 0 for $s > t$.

Theorem A47 (Ito Integral as a Continuous Martingale) *For any f in $\mathcal{H}^2$, there exists a continuous martingale X_t adapted to the standard Brownian filtration H_t such that*

$$X_t = \int_0^T f(\omega, s) 1(s \le t) dW(s) \quad \text{for all } t \le T$$

This continuous martingale we will denote by $\int_0^t f \, dW$.

So far we have defined a stochastic integral only for functions f that are square integral, that is, satisfy $E[\int_0^T f^2(\omega, t)dt] < \infty$, but this condition is too restrictive for some applications. A larger class of functions to which we can extend the notion of integral is the set of locally square integrable functions, $\mathcal{L}_{LOC}^2$. The word "local" in martingale and stochastic integration theory is a bit of a misnomer. A property holds locally if there is a sequence of stopping times ν_n each of which is finite but the $\nu_n \to \infty$, and the property holds when restricted to times $t \le \nu_n$.

Definition Let $\mathcal{L}_{LOC}^2$ be the set of functions $f(\omega, t)$ on the product space $\Omega \times [0, T]$ such that

1. f is measurable with respect to the product sigma field on $\Omega \times [0, T]$.
2. For each $t \in [0, T]$, $f(., t)$ is H_t-measurable (in other words, the stochastic process $f(., t)$ is adapted to H_t).

3. $P(\int_0^T f^2(\omega, s)ds < \infty) = 1.$

Clearly, this space includes $\mathcal{H}^2$ and arbitrary continuous functions of a Brownian motion. For any function f in $\mathcal{L}^2_{LOC}$, it is possible to define a sequence of stopping times

$$\nu_n = \min\left(T, \inf\left\{s; \int_0^s f^2(\omega, t)dt \geq n\right\}\right)$$

that acts as a *localizing* sequence for f. Such a sequence has the properties

1. ν_n is a nondecreasing sequence of stopping times.
2. $P[\nu_n = T \text{ for some } n] = 1.$
3. The functions $f_n(\omega, t) = f(\omega, t)1(t \leq \nu_n) \in \mathcal{H}^2$ for each n.

The purpose of the localizing sequence is essentially to provide approximations of a function f in $\mathcal{L}^2_{LOC}$ with functions $f(\omega, t)1(t \leq \nu_n)$ that are in $\mathcal{H}^2$ and therefore have a well-defined Ito integral as described above. The integral of f is obtained by taking the limit as $n \to \infty$ of the functions $f(\omega, t)1(t \leq \nu_n)$:

$$\int_0^t f(\omega, s)dW_s = \lim_{n\to\infty} \int_0^t f(\omega, t)1(t \leq \nu_n)dW_s$$

If f happens to be a continuous **nonrandom** function on $[0, T]$, the integral $\int_0^T f(s)dW_s$ is a limit in probability of the Riemann sums,

$$\sum f(s_i)(W_{t_{i+1}} - W_{t_i})$$

for any $t_i \leq s_i \leq t_{t+1}$. The integral is the limit of sums of the independent normal zero-mean random variables $f(s_i)(W_{t_{i+1}} - W_{t_i})$ and is therefore normally distributed. In fact,

$$X_t = \int_0^t f(s)dW_s$$

is a zero-mean Gaussian process with $\text{cov}(X_s, X_t) = \int_0^{\min(s,t)} f^2(u)du$. Such Gaussian processes are essentially time-changed Brownian motion processes according to the following.

Theorem A48 (Time Change to Brownian Motion) *Suppose $f(s)$ is a continuous nonrandom function on $[0, \infty)$ such that*

$$\int_0^\infty f^2(s)ds = \infty$$

Define the function $t(u) = \int_0^u f^2(s)ds$ *and its inverse function* $\tau(t) = \inf\{u; t(u) \geq t\}$. *Then*

$$Y(t) = \int_0^{\mathcal{T}(t)} f(s)dW_s$$

is a standard Brownian motion.

Definition: Local Martingale The process $M(t)$ is a local martingale with respect to the filtration H_t if there exists a nondecreasing sequence of stopping times $\tau_k \to \infty$ *a.s.* such that the processes

$$M_t^{(k)} = M(\min(t, \tau_k)) - M(0)$$

are martingales with respect to the same filtration.

In general, for $f \in \mathcal{L}^2_{LOC}$, stochastic integrals are local martingales, or more formally, there is a continuous local martingale equal (with probability 1) to the stochastic integral $\int_0^t f(\omega, s)dW_s$ for all t. We do not usually distinguish among processes that differ on a set of probability zero, so we assume that $\int_0^t f(\omega, s)dW_s$ is a continuous local martingale. There is a famous converse to this result, the martingale representation theorem, which asserts that a martingale can be written as a stochastic integral. We assume that H_t is the standard filtration of a Brownian motion W_t.

Theorem A49 (Martingale Representation Theorem) *Let* X_t *be an* H_t *martingale with* $E(X_T^2) < \infty$. *Then there exists* $\phi \in \mathcal{H}^2$ *such that*

$$X_t = X_0 + \int_0^t \phi(\omega, s)dW_s \quad \textit{for } 0 < t < T$$

and this representation is unique.

3.4 DIFFERENTIAL NOTATION AND ITO'S FORMULA

Summary 1: Rules of Box Algebra

It is common to use differential notation for stochastic differential equations such as

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t$$

to indicate (this is its only possible meaning) a stochastic process X_t, which is a solution of the equation written in integral form:

$$X_t = X_0 + \int_0^t \mu(s, X_s)ds + \int_0^t \mu(s, Xs)dW_s$$

We assume that the functions μ and σ are such that these two integrals, one a regular Riemann integral and the other a stochastic integral, are well defined, and we would like conditions on μ, σ such that existence and uniqueness of a solution is guaranteed. The following result is a standard one in this direction.

Theorem A50 (Existence and Uniqueness of Solutions of SDE) *Consider the stochastic DE*

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t \tag{3.8}$$

with initial condition $X_0 = x_0$. *Suppose for all* $0 < t < T$,

$$|\mu(t,x) - \mu(t,y)|^2 + |\sigma(t,x) - \sigma(t,y)|^2 \leq K|x-y|^2$$

and

$$|\mu(t,x)|^2 + |\sigma(t,x)|^2 \leq K(1 + |x|^2)$$

Then there is a unique (with probability 1) continuous adapted solution to (3.8) *and it satisfies*

$$\sup_{0<t<T} E(X_t^2) < \infty.$$

It is not difficult to show that some condition is required in the above theorem to ensure that the solution is unique. For example, if we consider the purely deterministic equation $dX_t = 3X_t^{2/3}dt$ with initial condition $X(0) = 0$, it has possible solutions $X_t = 0, t \leq a$, and $X_t = (t-a)^3, t > a$, for arbitrary $a > 0$. There are at least as many distinct solutions as there are possible values of a.

Now suppose a process X_t is a solution of (3.8) and we are interested an a new stochastic process defined as a function of X_t, say $Y_t = f(t, X_t)$. Ito's formula is used to write Y_t with a stochastic differential equation similar to (3.8). Suppose we attempt this using a Taylor series expansion where we will temporarily regard differentials such as dt, dX_t as small increments of time and the process, respectively (notation such as Δt, ΔW might have been preferable here). Let the partial derivatives of f be denoted by

$$f_1(t,x) = \frac{\partial f}{\partial t}, \quad f_2(t,x) = \frac{\partial f}{\partial x}, \quad f_{22}(t,x) = \frac{\partial^2 f}{\partial x^2}$$

and so on. Then the Taylor series expansion can be written

$$\begin{aligned} dY_t =& f_1(t,X_t)dt + \frac{1}{2}f_{11}(t,X_t)(dt)^2 + \cdots \\ &+ f_2(t,X_t)dX_t + \frac{1}{2}f_{22}(t,X_t)(dX_t)^2 + \cdots \\ &+ f_{12}(t,X_t)(dt)(dX_t) + \cdots \end{aligned} \tag{3.9}$$

and although there are infinitely many terms in this expansion, all but a few turn out to be negligible. The contribution of these terms is largely determined by some simple rules, often referred to as the rules of box algebra. In an expansion to terms of order dt, as $dt \to 0$ higher-order terms such as $(dt)^j$ are all negligible for $j > 0$. For example, $(dt)^2 = o(dt)$ as $dt \to 0$ (intuitively this means that $(dt)^2$ goes to zero faster than dt does). Similarly, cross terms such as $(dt)(dW_t)$ are negligible because the increment dW_t is normally distributed with mean 0 and standard deviation $(dt)^{1/2}$, and so $(dt)(dW_t)$ has standard deviation $(dt)^{3/2} = o(dt)$. We summarize some of these order arguments with the oversimplified rules below, where the symbol $\sim$ is taken to mean "is order of, as $dt \to 0$."

$$\begin{aligned} (dt)(dt) &\sim 0 \\ (dt)(dW_t) &\sim 0 \\ (dW_t)(dW_t) &\sim dt \end{aligned}$$

From these we can obtain, for example,

$$\begin{aligned} (dX_t)(dX_t) &= [\mu(t,X_t)dt + \sigma(t,X_t)dW_t][\mu(t,X_t)dt + \sigma(t,X_t)dW_t] \\ &= \mu^2(t,X_t)(dt)^2 + 2\mu(t,X_t)\sigma(t,X_t)(dt)(dW_t) \\ &\quad + \sigma^2(t,X_t)(dW_t)(dW_t) \sim \sigma^2(t,X_t)dt \end{aligned}$$

which indicates the order of the small increments in the process X_t. If we now use these rules to evaluate (3.9), we obtain

$$\begin{aligned} dY_t &\sim f_1(t,X_t)dt + f_2(t,X_t)dX_t + \frac{1}{2}f_{22}(t,X_t)(dX_t)^2 \\ &\sim f_1(t,X_t)dt + f_2(t,X_t)(\mu(t,X_t)dt + \sigma(t,X_t)dW_t) \\ &\quad + \frac{1}{2}f_{22}(t,X_t)\sigma^2(t,X_t)dt \end{aligned}$$

which is the differential expression of Ito's formula.

Theorem A51 (Ito's Formula) *Suppose X_t satisfies $dX_t = \mu(t,X_t)dt + \sigma(t,X_t)dW_t$. Then for any function f such that f_1 and f_{22} are continuous, the process $f(t,X_t)$ satisfies the stochastic differential equation*

$$\begin{aligned} df(t,X_t) &= \{\mu(t,X_t)f_2(t,X_t) + f_1(t,X_t) + \frac{1}{2}f_{22}(t,X_t)\sigma^2(t,X_t)\}dt \\ &\quad + f_2(t,X_t)\sigma(t,X_t)dW_t \end{aligned}$$

Example: Geometric Brownian Motion Suppose X_t satisfies

$$dX_t = aX_t\, dt + \sigma X_t\, dW_t$$

and $f(t, X_t) = \ln(X_t)$. Then substituting in Ito's formula, since $f_1 = 0, f_2 = X_t^{-1}, f_{22} = -X_t^{-2}$,

$$\begin{aligned} dY_t &= X_t^{-1} a X_t\, dt - \frac{1}{2} X_t^{-2} \sigma^2 X_t^2\, dt + X_t^{-1} \sigma X_t\, dW_t \\ &= \left(a - \frac{\sigma^2}{2}\right) dt + \sigma\, dW_t \end{aligned}$$

and so $Y_t = \ln(X_t)$ is a Brownian motion with drift $a - \frac{\sigma^2}{2}$ and volatility σ.

Example: Ornstein-Uhlenbeck Process Consider the stochastic process defined as

$$X_t = x_0 e^{-\alpha t} + \sigma e^{-\alpha t} \int_0^t e^{\alpha s}\, dW_s$$

for parameters $\alpha, \sigma > 0$. Then,

$$\begin{aligned} dX_t &= (-\alpha) x_0 e^{-\alpha t} dt + (-\alpha) \sigma e^{-\alpha t} \int_0^t e^{\alpha s}\, dW_s + \sigma e^{-\alpha t} e^{\alpha t}\, dW_t \\ &= -\alpha X_t\, dt + \sigma\, dW_t. \end{aligned}$$

with the initial condition $X_0 = x_0$. This process has Gaussian increments and covariance structure $\text{cov}(X_s, X_t) = \sigma^2 \int_0^s e^{-\alpha(s+t-u)} ds$, for $s < t$, and is called the Ornstein-Uhlenbeck process.

Example: Brownian Bridge Consider the process defined as

$$X_t = (1 - t) \int_0^t \frac{1}{1 - s} dW_s, \qquad \text{for } 0 < t < 1$$

subject to the initial condition $X_0 = 0$. Then

$$\begin{aligned} dX_t &= -\int_0^t \frac{1}{1 - s} dW_s + (1 - t) \frac{1}{1 - t} dW_t \\ &= -\frac{X_t}{1 - t} dt + dW_t \end{aligned}$$

This process satisfying $X_0 = X_1 = 0$ and

$$dX_t = -\frac{X_t}{1 - t} dt + dW_t$$

is called the Brownian bridge. It can also be constructed as $X_t = W_t - t W_1$. The distribution of the Brownian bridge is identical to the conditional distribution of a standard Brownian motion W_t given that $W_0 = 0$ and $W_1 = 0$. The Brownian bridge is a Gaussian process with covariance $\text{cov}(X_s, X_t) = s(1 - t)$ for $s < t$.

Theorem A52 (Ito's Formula for Two Processes) *If*

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t$$
$$dY_t = \alpha(t, Y_t)dt + \beta(t, Y_t)dW_t$$

then

$$\begin{aligned} df(X_t, Y_t) =& f_1(X_t, Y_t)dX_t + f_2(X_t, Y_t)dY_t \\ &+ \frac{1}{2}f_{11}(X_t, Y_t)b^2\,dt + \frac{1}{2}f_{22}(X_t, Y_t)\beta^2\,dt \\ &+ f_{12}(X_t, Y_t)b\beta\,dt \end{aligned}$$

There is an immediate application of this result to obtain the product rule for differentiation of diffusion processes. If we put $f(x, y) = xy$ above, we obtain

$$d(X_tY_t) = Y_t\,dX_t + X_t\,dY_t + b\beta\,dt$$

This product rule reduces to the usual with either β or b identically 0.

3.5 QUADRATIC VARIATION

One way of defining the variation of a process X_t is to choose a partition $\pi = \{0 = t_0 \le t_1 \le \cdots \le t_n = t\}$ and then define $Q_\pi(X_t) = \sum_i (X_{t_i} - X_{t_{i-1}})^2$.

For a diffusion process

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t$$

satisfying standard conditions, as the mesh size $\max|t_i - t_{i-1}|$ converges to zero, we have $Q_\pi(X_t) \to \int_0^t \sigma^2(s, X_s)ds$ in probability. This limit $\int_0^t \sigma^2(s, X_s)ds$ is the process that we earlier denoted $<X, X>_t$. For brevity, the redundancy in the notation is usually removed, and the process $<X, X>_t$ is denoted $<X>_t$. For diffusion processes, variation of lower order such as $\sum_i |X_{t_i} - X_{t_{i-1}}|$ approaches infinity and variation of higher order, for example, $\sum_i (X_{t_i} - X_{t_{i-1}})^4$, converges to zero as the mesh size converges. We will return to the definition of the predictable covariation process $<X, Y>_t$ in a more general setting shortly.

The Stochastic Exponential Suppose X_t is a diffusion process and consider a stochastic differential equation

$$dY_t = Y_t\,dX_t \tag{3.10}$$

with initial condition $Y_0 = 1$. If X_t were an ordinary differentiable function, we could solve this equation by integrating both sides of

$$\frac{dY_t}{Y_t} = dX_t$$

to obtain the exponential function

$$Y_t = c\exp(X_t) \tag{3.11}$$

where c is a constant of integration. We might try and work backward from (3.11) to see if this is the correct solution in the general case in which X_t is a diffusion. Letting $f(X_t) = \exp(X_t)$ and using Ito's formula,

$$\begin{aligned} df(X_t) &= \left\{\exp(X_t) + \frac{1}{2}\exp(X_t)\sigma^2(t, X_t)\right\}dt + \exp(X_t)\sigma(t, X_t)dW_t \\ &\neq f(X_t)dX_t \end{aligned}$$

so this solution is not quite right. There is, however, a minor fix of the exponential function that does provide a solution. Suppose we try a solution of the form

$$Y_t = f(t, X_t) = \exp(X_t + h(t))$$

where $h(t)$ is some differentiable stochastic process. Then again using Ito's lemma, since $f_1(t, X_t) = Y_t h'(t)$ and $f_2(t, X_t) = f_{22}(t, X_t) = Y_t$,

$$\begin{aligned} dY_t &= f_1(t, X_t)dt + f_2(t, X_t)dX_t - \frac{1}{2}f_{22}(t, X_t)\sigma^2(t, X_t)dt \\ &= Y_t\{h'(t) + \mu(t, X_t) + \frac{1}{2}\sigma^2(t, X_t)\}dt + Y_t\sigma(t, X_t))dW_t \end{aligned}$$

and if we choose just the right function h so that $h'(t) = -\frac{1}{2}\sigma^2(t, X_t)$, we can get a solution to (3.10). Since $h(t) = -\frac{1}{2}\int_0^t \sigma^2(s, X_s)ds$ the solution is

$$Y_t = \exp\left(X_t - \frac{1}{2}\int_0^t \sigma^2(s, X_s)ds\right) = \exp\left\{X_t - \frac{1}{2} < X >_t\right\}$$

We may denote this solution $Y = \mathcal{E}(X)$. We saw earlier that $\mathcal{E}(\alpha W)$ is a martingale for W a standard Brownian motion and α real. Since the solution to this equation is an exponential in the ordinary calculus, the term "stochastic exponential" seems justified. The "extra" term in the exponent $\frac{1}{2} < X >_t$ is a consequence of the infinite local variation of the process X_t. One of the most common conditions for $\mathcal{E}(X)$ to be a martingale is the following:

Novikov's Condition Suppose that for $g \in \mathcal{L}^2_{LOC}$

$$E\exp\left\{\frac{1}{2}\int_0^T g^2(w, s_s)ds\right\} < \infty$$

Then $M_t = \mathcal{E}(\int_0^t g(\omega, s)dW_s)$ is a martingale.

3.6 SEMIMARTINGALES

Suppose M_t is a continuous martingale adapted to a filtration H_t, and A_t is a continuous adapted process that is nondecreasing. It is easy to see that the sum $A_t + M_t$ is a submartingale. But can this argument be reversed? If we are given a submartingale X_t, is it possible to find a nondecreasing process A_t and a martingale M_t such that $X_t = A_t + M_t$? The fundamental result in this direction is the Doob-Meyer decomposition.

Theorem A53 (Doob-Meyer Decomposition) *Let X be a continuous submartingale adapted to a filtration H_t. Then X can be uniquely written as $X_t = A_t + M_t$, where M_t is a local martingale and A_t is an adapted nondecreasing process such that $A_0 = 0$.*

Recall that if M_t is a square integrable martingale, then M_t^2 is a submartingale (this follows from Jensen's inequality). Then according to the Doob-Meyer decomposition, we can decompose M_t^2 into two components, one a martingale and the other a nondecreasing continuous adapted process, which we call the (predictable) *quadratic variation process* $<M>_t$. In other words,

$$M_t^2 - <M>_t$$

is a continuous martingale. We may take this as the the more general definition of the process $<M>$, encountered earlier for processes obtained as stochastic integrals. For example, suppose

$$X_t(\omega) = \int_0^t f(\omega, t) dW(t)$$

where $f \in \mathcal{H}^2$. Then with $<X>_t = \int_0^t f^2(\omega, t)dt$ and $M_t = X_t^2 - <X>_t$, notice that for $s < t$

$$\begin{aligned} E[M_t - M_s | H_s] &= E\left[\left\{ \int_s^t f(\omega, u) dW(u) \right\}^2 - \int_s^t f^2(\omega, u) du | H_s \right] \\ &= 0 \end{aligned}$$

by (3.5). This means that our earlier definition of the process $<X>$ coincides with the current one. For two martingales X, Y, we can define the predictable covariation process $<X, Y>$ as

$$<X, Y>_t = \frac{1}{4}\{<X + Y>_t - <X - Y>_t\}$$

and once again this agrees with the earlier definition in Section 3.3, since if

X and Y are defined as

$$X_t(\omega) = \int_0^t f(\omega, t)dW(t)$$
$$Y_t(\omega) = \int_0^t g(\omega, t)dW(t)$$

the predictable covariation is

$$< X, Y >_t (\omega) = \int_0^t f(\omega, t)g(\omega, t)dt$$

This also follows from the Ito isometry.

Definition: Semimartingale A continuous adapted process X_t is a *semimartingale* if it can be written as the sum $X_t = A_t + M_t$ of a continuous adapted process A_t of locally bounded variation, and a continuous local martingale M_t.

The stochastic integral for square integrable martingales can be extended to the class of semimartingales. Let $X_t = A_t + M_t$ be a continuous semimartingale. We define

$$\int h(t)dX_t = \int h(t)dA_t + \int h(t)dM_t \tag{3.12}$$

The first integral on the right-hand side of (3.12) is understood to be a Lebesgue-Stieltjes integral while the second is an Ito stochastic integral. There are a number of details that need to be checked with this definition — for example, whether when we decompose a semimartingale into the two components, one with bounded variation and one a local martingale in two different ways (this decomposition is not unique), the same integral is obtained.

3.7 GIRSANOV'S THEOREM

Consider the Brownian motion defined by

$$dX_t = \mu\, dt + dW_t$$

with μ a constant drift parameter and denote by $E_\mu(.)$ the expectation when the drift is μ. Let $f_\mu(x)$ be the $N(\mu, T)$ probability density function. Then we can compute expectations under nonzero drift μ using a Brownian motion that has drift zero since

$$E_\mu(g(X_T)) = E_0\{g(X_T)M_T(X)\}$$

where

$$M_t(X) = \mathcal{E}(\mu X) = \exp\left\{\mu X_t - \frac{1}{2}\mu^2 t\right\}.$$

This is easy to check since the stochastic exponential $M_T(X)$ happens to be the ratio of the $N(\mu T, T)$ probability density function to the $N(0, T)$ density. The implications are many and useful. We can, for example, calculate moments or simulate under the condition $\mu = 0$ and apply the results to the case $\mu \neq 0$. By a similar calculation, for a bounded Borel measurable function $g(X_{t_1}, \ldots, X_{t_n})$, where $0 \le t_1 \le \cdots \le t_n$,

$$E_\mu\{g(X_{t_1}, \ldots, X_{t_n})\} = E_0\{g(X_{t_1}, \ldots, X_{t_n})M_{t_n}(X)\}$$

Theorem A54 (Girsanov's Theorem for Brownian Motion) *Consider a Brownian motion with drift μ defined by*

$$X_t = \mu t + W_t$$

Then for any bounded measurable function g defined on the space $C[0, T]$ of the paths, we have

$$E_\mu[g(X)] = E_0[g(X)M_T(X)]$$

where again $M_T(X)$ is the exponential martingale $\mathcal{E}(\mu X) = \exp\big(\mu X_T - \frac{1}{2}\mu^2 T\big)$.

Note that if we let P_0, P_μ denote the measures on the function space corresponding to drift 0 and μ, respectively, we can formally write

$$\begin{aligned} E_\mu[g(X)] &= \int g(x)dP_\mu = \int g(x)\frac{dP_\mu}{dP_0}dP_0 \\ &= E_0\left\{g(X)\frac{dP_\mu}{dP_0}\right\} \end{aligned}$$

which means that $M_T(X)$ plays the role of a likelihood ratio,

$$\frac{dP_\mu}{dP_0}$$

for a restriction of the process to the interval $[0, T]$. If $g(X)$ depended only on the process up to time $t < T$, then, from the martingale property of $M_t(X)$,

$$\begin{aligned} E_\mu[g(X)] &= E_0[g(X)M_T(X)] \\ &= E_0\{E[g(X)M_T(X)|H_t]\} \\ &= E_0\{g(X)M_t(X)\} \end{aligned}$$

which shows that $M_t(X)$ plays the role of a likelihood ratio for a restriction of the process to the interval $[0, t]$.

We can argue for the form of $M_t(X)$ and show that it "should" be a martingale under $\mu = 0$ by considering the limit of the ratio of the finite-dimensional probability density functions such as

$$\frac{f_\mu(x_{t_1}, ..., x_{t_n})}{f_0(x_{t_1}, ..., x_{t_n})}$$

where f_μ denotes the joint probability density function of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ for $t_1 < t_2 < \cdots < t_n = T$. These likelihood ratios are discrete-time martingales under P_0. For a more general diffusion, provided that the diffusion terms are identical, we can still express the Radon-Nikodym derivative as a stochastic exponential.

Theorem A55 (Girsanov's Theorem) *Suppose P is the measure on $C[0, T]$ induced by $X_0 = 0$, and*

$$dX_t = \mu(\omega, t)dt + \sigma(\omega, t)dW_t$$

under P. Assume the standard conditions so that the corresponding stochastic integrals are well defined. Assume that the function

$$\theta(\omega, t) = \frac{\mu(\omega, t) - \nu(\omega, t)}{\sigma(\omega, t)}$$

is bounded. Then the stochastic exponential

$$M_t = \mathcal{E}\left(-\int_0^t \theta(\omega, s)dW_s\right) = \exp\left\{-\int_0^t \theta(\omega, s)dW_s - \frac{1}{2}\int_0^t \theta^2(\omega, s)ds\right\}$$

is a martingale under P. Suppose we define a measure Q on $C[0, T]$ by

$$\frac{dQ}{dP} = M_T$$

or, equivalently, for measurable subsets A,

$$Q(A) = E_P[1(A)M_T]$$

Then under the measure Q, the process W_t' defined by

$$W_t' = W_t - \int_0^t \theta(\omega, s)dW_s$$

is a standard Brownian motion, and X_t has the representation

$$dX_t = \nu(\omega, t)dt + \sigma(\omega, t)dW_t'$$