STOCHASTIC PROCESSES

Second Edition

Sheldon M. Ross
University of California, Berkeley

JOHN WILEY & SONS, INC.
New York • Chichester • Brisbane • Toronto • Singapore

CHAPTER 1

Preliminaries

1.1 Probability

A basic notion in probability theory is *random experiment:* an experiment whose outcome cannot be determined in advance. The set of all possible outcomes of an experiment is called the *sample space* of that experiment, and we denote it by S.

An *event* is a subset of a sample space, and is said to occur if the outcome of the experiment is an element of that subset. We shall suppose that for each event E of the sample space S a number $P(E)$ is defined and satisfies the following three axioms*:

Axiom (1) $0 \leq P(E) \leq 1$.

Axiom (2) $P(S) = 1$.

Axiom (3) For any sequence of events $E_1, E_2, \ldots$ that are mutually exclusive, that is, events for which $E_iE_j = \phi$ when $i \neq j$ (where ϕ is the null set),

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

We refer to $P(E)$ as the probability of the event E.

Some simple consequences of axioms (1), (2), and (3) are:

1.1.1. If $E \subset F$, then $P(E) \leq P(F)$.

1.1.2. $P(E^c) = 1 - P(E)$ where E^c is the complement of E.

1.1.3. $P(\bigcup_1^n E_i) = \sum_1^n P(E_i)$ when the E_i are mutually exclusive.

1.1.4. $P(\bigcup_1^\infty E_i) \leq \sum_1^\infty P(E_i)$.

The inequality (1.1.4) is known as *Boole's inequality.*

* Actually $P(E)$ will only be defined for the so-called measurable events of S. But this restriction need not concern us.

An important property of the probability function P is that it is continuous. To make this more precise, we need the concept of a limiting event, which we define as follows: A sequence of events $\{E_n, n \geq 1\}$ is said to be an *increasing* sequence if $E_n \subset E_{n+1}$, $n \geq 1$ and is said to be *decreasing* if $E_n \supset E_{n+1}$, $n \geq 1$. If $\{E_n, n \geq 1\}$ is an increasing sequence of events, then we define a new event, denoted by $\lim_{n\to\infty} E_n$ by

$$\lim_{n\to\infty} E_n = \bigcup_{i=1}^{\infty} E_i \qquad \text{when } E_n \subset E_{n+1},\ n \geq 1.$$

Similarly if $\{E_n, n \geq 1\}$ is a decreasing sequence, then define $\lim_{n\to\infty} E_n$ by

$$\lim_{n\to\infty} E_n = \bigcap_{i=1}^{\infty} E_i, \qquad \text{when } E_n \supset E_{n+1},\ n \geq 1.$$

We may now state the following:

PROPOSITION 1.1.1

If $\{E_n, n \geq 1\}$ is either an increasing or decreasing sequence of events, then

$$\lim_{n\to\infty} P(E_n) = P\left(\lim_{n\to\infty} E_n\right).$$

Proof Suppose, first, that $\{E_n, n \geq 1\}$ is an increasing sequence, and define events F_n, $n \geq 1$ by

$$F_1 = E_1,$$

$$F_n = E_n\left(\bigcup_{1}^{n-1} E_i\right)^c = E_n E_{n-1}^c, \qquad n > 1.$$

That is, F_n consists of those points in E_n that are not in any of the earlier E_i, $i < n$. It is easy to verify that the F_n are mutually exclusive events such that

$$\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} E_i \quad \text{and} \quad \bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i \qquad \text{for all } n \geq 1.$$

Thus

$$
\begin{aligned}
P\left(\bigcup_1^\infty E_i\right) &= P\left(\bigcup_i^\infty F_i\right) \\
&= \sum_1^\infty P(F_i) \qquad \text{(by Axiom 3)} \\
&= \lim_{n\to\infty} \sum_1^n P(F_i) \\
&= \lim_{n\to\infty} P\left(\bigcup_1^n F_i\right) \\
&= \lim_{n\to\infty} P\left(\bigcup_1^n E_i\right) \\
&= \lim_{n\to\infty} P(E_n),
\end{aligned}
$$

which proves the result when $\{E_n, n \geq 1\}$ is increasing.

If $\{E_n, n \geq 1\}$ is a decreasing sequence, then $\{E_n^c, n \geq 1\}$ is an increasing sequence; hence,

$$P\left(\bigcup_1^\infty E_n^c\right) = \lim_{n\to\infty} P(E_n^c).$$

But, as $\bigcup_1^\infty E_n^c = (\bigcap_1^\infty E_n)^c$, we see that

$$1 - P\left(\bigcap_1^\infty E_n\right) = \lim_{n\to\infty} [1 - P(E_n)],$$

or, equivalently,

$$P\left(\bigcap_1^\infty E_n\right) = \lim_{n\to\infty} P(E_n),$$

which proves the result.

EXAMPLE 1.1(A) Consider a population consisting of individuals able to produce offspring of the same kind. The number of individuals

initially present, denoted by X_0, is called the size of the zeroth generation. All offspring of the zeroth generation constitute the first generation and their number is denoted by X_1. In general, let X_n denote the size of the nth generation.

Since $X_n = 0$ implies that $X_{n+1} = 0$, it follows that $P\{X_n = 0\}$ is increasing and thus $\lim_{n\to\infty} P\{X_n = 0\}$ exists. What does it represent? To answer this use Proposition 1.1.1 as follows:

$$\begin{aligned}\lim_{n\to\infty} P\{X_n = 0\} &= P\left\{\lim_{n\to\infty}\{X_n = 0\}\right\} \\ &= P\left\{\bigcup_n \{X_n = 0\}\right\} \\ &= P\{\text{the population ever dies out}\}.\end{aligned}$$

That is, the limiting probability that the nth generation is void of individuals is equal to the probability of eventual extinction of the population.

Proposition 1.1.1 can also be used to prove the Borel-Cantelli lemma.

PROPOSITION 1.1.2

The Borel-Cantelli Lemma
Let $E_1, E_2, \ldots$ denote a sequence of events. If

$$\sum_{i=1}^{\infty} P(E_i) < \infty,$$

then

$$P\{\text{an infinite number of the } E_i \text{ occur}\} = 0.$$

Proof The event that an infinite number of the E_i occur, called the $\limsup_{i\to\infty} E_i$, can be expressed as

$$\limsup_{i\to\infty} E_i = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} E_i.$$

This follows since if an infinite number of the E_i occur, then $\bigcup_{i=n}^{\infty} E_i$ occurs for each n and thus $\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} E_i$ occurs. On the other hand, if $\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} E_i$ occurs, then $\bigcup_{i=n}^{\infty} E_i$ occurs for each n, and thus for each n at least one of the E_i occurs where $i \geq n$; and, hence, an infinite number of the E_i occur.

As $\bigcup_{i=n}^{\infty} E_i$, $n \geq 1$, is a decreasing sequence of events, it follows from Proposition 1.1.1 that

$$\begin{aligned} P\left(\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} E_i\right) &= P\left(\lim_{n\to\infty} \bigcup_{i=n}^{\infty} E_i\right) \\ &= \lim_{n\to\infty} P\left(\bigcup_{i=n}^{\infty} E_i\right) \\ &\leq \lim_{n\to\infty} \sum_{i=n}^{\infty} P(E_i) \\ &= 0, \end{aligned}$$

and the result is proven.

a.

EXAMPLE 1.1(B) Let $X_1, X_2, \ldots$ be such that

$$P\{X_n = 0\} = 1/n^2 = 1 - P\{X_n = 1\}, \qquad n \geq 1.$$

If we let $E_n = \{X_n = 0\}$, then, as $\sum_n^{\infty} P(E_n) < \infty$, it follows from the Borel-Cantelli lemma that the probability that X_n equals 0 for an infinite number of n is equal to 0. Hence, for all n sufficiently large, X_n must equal 1, and so we may conclude that, with probability 1,

$$\lim_{n\to\infty} X_n = 1.$$

For a converse to the Borel-Cantelli lemma, independence is required.

$_i$, can

PROPOSITION 1.1.3

Converse to the Borel-Cantelli Lemma
If $E_1, E_2, \ldots$ are independent events such that

$$\sum_{n=1}^{\infty} P(E_n) = \infty,$$

r each
, then
re $i \geq$

then

$$P\{\text{an infinite number of the } E_n \text{ occur}\} = 1.$$

Proof

$$P\{\text{an infinite number of the } E_n \text{ occur}\} = P\left\{\lim_{n\to\infty} \bigcup_{i=n}^{\infty} E_i\right\}$$

$$= \lim_{n\to\infty} P\left(\bigcup_{i=n}^{\infty} E_i\right)$$

$$= \lim_{n\to\infty} \left[1 - P\left(\bigcap_{i=n}^{\infty} E_i^c\right)\right].$$

Now,

$$P\left(\bigcap_{i=n}^{\infty} E_i^c\right) = \prod_{i=n}^{\infty} P(E_i^c) \qquad \text{(by independence)}$$

$$= \prod_{i=n}^{\infty} (1 - P(E_i))$$

$$\leq \prod_{i=n}^{\infty} e^{-P(E_i)} \qquad \text{(by the inequality } 1 - x \leq e^{-x}\text{)}$$

$$= \exp\left(-\sum_{n}^{\infty} P(E_i)\right)$$

$$= 0 \quad \text{since } \sum_{i=n}^{\infty} P(E_i) = \infty \text{ for all } n.$$

Hence the result follows.

EXAMPLE 1.1(c) Let $X_1, X_2, \ldots$ be independent and such that

$$P\{X_n = 0\} = 1/n = 1 - P\{X_n = 1\}, \qquad n \geq 1.$$

If we let $E_n = \{X_n = 0\}$, then as $\sum_{n=1}^{\infty} P(E_n) = \infty$ it follows from Proposition 1.1.3 that E_n occurs infinitely often. Also, as $\sum_{n=1}^{\infty} P(E_n^c) = \infty$ it also follows that E_n^c also occurs infinitely often. Hence, with probability 1, X_n will equal 0 infinitely often and will also equal 1 infinitely often. Hence, with probability 1, X_n will not approach a limiting value as $n \to \infty$.

1.2 Random Variables

Consider a random experiment having sample space S. A *random variable* X is a function that assigns a real value to each outcome in S. For any set of real numbers A, the probability that X will assume a value that is contained in the set A is equal to the probability that the outcome of the experiment is contained in $X^{-1}(A)$. That is,

$$P\{X \in A\} = P(X^{-1}(A)),$$

where $X^{-1}(A)$ is the event consisting of all points $s \in S$ such that $X(s) \in A$.

The *distribution function* F of the random variable X is defined for any real number x by

$$F(x) = P\{X \leq x\} = P\{X \in (-\infty, x]\}.$$

We shall denote $1 - F(x)$ by $\bar{F}(x)$, and so

$$\bar{F}(x) = P\{X > x\}.$$

A random variable X is said to be *discrete* if its set of possible values is countable. For discrete random variables,

$$F(x) = \sum_{y \leq x} P\{X = y\}.$$

A random variable is called *continuous* if there exists a function $f(x)$, called the *probability density function*, such that

$$P\{X \text{ is in } B\} = \int_B f(x)\, dx$$

for every set B. Since $F(x) = \int_{-\infty}^{x} f(x)\, dx$, it follows that

$$f(x) = \frac{d}{dx} F(x).$$

The *joint distribution function* F of two random variables X and Y is defined by

$$F(x, y) = P\{X \leq x, Y \leq y\}.$$

The distribution functions of X and Y,

$$F_X(x) = P\{X \leq x\} \quad \text{and} \quad F_Y(y) = P\{Y \leq y\},$$

can be obtained from $F(x, y)$ by making use of the continuity property of the probability operator. Specifically, let y_n, $n \geq 1$, denote an increasing sequence converging to ∞. Then as the events $\{X \leq x, Y \leq y_n\}$, $n \geq 1$, are increasing and

$$\lim_{n\to\infty}\{X \leq x, Y \leq y_n\} = \bigcup_{n=1}^{\infty} \{X \leq x, Y \leq y_n\} = \{X \leq x\},$$

it follows from the continuity property that

$$\lim_{n\to\infty} P\{X \leq x, Y \leq y_n\} = P\{X \leq x\},$$

or, equivalently,

$$F_X(x) = \lim_{y\to\infty} F(x, y).$$

Similarly,

$$F_Y(y) = \lim_{x\to\infty} F(x, y).$$

The random variables X and Y are said to be *independent* if

$$F(x, y) = F_X(x)F_Y(y)$$

for all x and y.

The random variables X and Y are said to be *jointly continuous* if there exists a function $f(x, y)$, called the *joint probability density function*, such that

$$P\{X \text{ is in } A, Y \text{ is in } B\} = \int_A \int_B f(x, y)\, dy\, dx$$

for all sets A and B.

The joint distribution of any collection $X_1, X_2, \ldots, X_n$ of random variables is defined by

$$F(x_1, \ldots, x_n) = P\{X_1 \leq x_1, \ldots, X_n \leq x_n\}.$$

Furthermore, the n random variables are said to be independent if

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2) \cdots F_{X_n}(x_n),$$

where

$$F_{X_i}(x_i) = \lim_{\substack{x_j\to\infty \\ j\neq i}} F(x_1, \ldots, x_n).$$

1.3 Expected Value

The *expectation* or *mean* of the random variable X, denoted by $E[X]$, is defined by

$$E[X] = \int_{-\infty}^{\infty} x\, dF(x) \tag{1.3.1}$$

$$= \begin{cases} \int_{-\infty}^{\infty} xf(x)\, dx & \text{if } X \text{ is continuous} \\ \sum_{x} xP\{X = x\} & \text{if } X \text{ is discrete} \end{cases}$$

provided the above integral exists.

Equation (1.3.1) also defines the expectation of any function of X, say $h(X)$. Since $h(X)$ is itself a random variable, it follows from (1.3.1) that

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

where F_h is the distribution function of $h(X)$. However, it can be shown that this is identical to $\int_{-\infty}^{\infty} h(x)\, dF(x)$. That is,

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)\, dF(x). \tag{1.3.2}$$

The variance of the random variable X is defined by

$$\begin{aligned} \operatorname{Var} X &= E[(X - E[X])^2] \\ &= E[X^2] - E^2[X]. \end{aligned}$$

Two jointly distributed random variables X and Y are said to be uncorrelated if their covariance, defined by

$$\begin{aligned} \operatorname{Cov}(X, Y) &= E[(X - EX)(Y - EY)] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

is zero. It follows that independent random variables are uncorrelated. However, the converse need not be true. (The reader should think of an example.)

An important property of expectations is that the expectation of a sum of random variables is equal to the sum of the expectations.

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]. \tag{1.3.3}$$

The corresponding property for variances is that

$$\text{(1.3.4)} \qquad \operatorname{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2\sum\sum_{i<j} \operatorname{Cov}(X_i, X_j).$$

EXAMPLE 1.3(A) ***The Matching Problem.*** At a party n people put their hats in the center of a room where the hats are mixed together. Each person then randomly selects one. We are interested in the mean and variance of X—the number that select their own hat.

To solve, we use the representation

$$X = X_1 + X_2 + \cdots + X_n,$$

where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th person selects his or her own hat} \\ 0 & \text{otherwise.} \end{cases}$$

Now, as the ith person is equally likely to select any of the n hats, it follows that $P\{X_i = 1\} = 1/n$, and so

$$E[X_i] = 1/n,$$

$$\operatorname{Var}(X_i) = \frac{1}{n}\left(1 - \frac{1}{n}\right) = \frac{n-1}{n^2}.$$

Also

$$\operatorname{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j].$$

Now,

$$X_i X_j = \begin{cases} 1 & \text{if the } i\text{th and } j\text{th party goers both select their own hats} \\ 0 & \text{otherwise,} \end{cases}$$

and thus

$$\begin{aligned} E[X_i X_j] &= P\{X_i = 1, X_j = 1\} \\ &= P\{X_i = 1\}P\{X_j = 1 \mid X_i = 1\} \\ &= \frac{1}{n}\frac{1}{n-1}. \end{aligned}$$

Hence,

$$\text{Cov}(X_i, X_j) = \frac{1}{n(n-1)} - \left(\frac{1}{n}\right)^2 = \frac{1}{n^2(n-1)}.$$

Therefore, from (1.3.3) and (1.3.4),

$$E[X] = 1$$

and

$$\begin{aligned}\text{Var}(X) &= \frac{n-1}{n} + 2\binom{n}{2}\frac{1}{n^2(n-1)} \\ &= 1.\end{aligned}$$

Thus both the mean and variance of the number of matches are equal to 1. (See Example 1.5(f) for an explanation as to why these results are not surprising.)

EXAMPLE 1.3(B) ***Some Probability Identities.*** Let $A_1, A_2, \ldots, A_n$ denote events and define the indicator variables $I_j, j = 1, \ldots, n$ by

$$I_j = \begin{cases} 1 & \text{if } A_j \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

Letting

$$N = \sum_{j=1}^{n} I_j,$$

then N denotes the number of the A_j, $1 \le j \le n$, that occur. A useful identity can be obtained by noting that

$$(1.3.5) \qquad (1-1)^N = \begin{cases} 1 & \text{if } N = 0 \\ 0 & \text{if } N > 0. \end{cases}$$

But by the binomial theorem,

$$(1.3.6) \qquad \begin{aligned}(1-1)^N &= \sum_{i=0}^{N} \binom{N}{i} (-1)^i \\ &= \sum_{i=0}^{n} \binom{N}{i} (-1)^i \qquad \text{since } \binom{m}{i} = 0 \text{ when } i > m.\end{aligned}$$

Hence, if we let

$$I = \begin{cases} 1 & \text{if } N > 0 \\ 0 & \text{if } N = 0, \end{cases}$$

then (1.3.5) and (1.3.6) yield

$$1 - I = \sum_{i=0}^{n} \binom{N}{i} (-1)^i$$

or

$$I = \sum_{i=1}^{n} \binom{N}{i} (-1)^{i+1}. \tag{1.3.7}$$

Taking expectations of both sides of (1.3.7) yields

$$E[I] = E[N] - E\left[\binom{N}{2}\right] + \cdots + (-1)^{n+1} E\left[\binom{N}{n}\right]. \tag{1.3.8}$$

However,

$$\begin{aligned} E[I] &= P\{N > 0\} \\ &= P\{\text{at least one of the } A_i \text{ occurs}\} \\ &= P\left(\bigcup_{1}^{n} A_i\right) \end{aligned}$$

and

$$E[N] = E\left[\sum_{j=1}^{n} I_j\right] = \sum_{j=1}^{n} P(A_j),$$

$$\begin{aligned} E\left[\binom{N}{2}\right] &= E[\text{number of pairs of the } A_i \text{ that occur}] \\ &= E\left[\sum\sum_{i<j} I_i I_j\right] \\ &= \sum\sum_{i<j} E[I_i I_j] \\ &= \sum\sum_{i<j} P(A_i A_j), \end{aligned}$$

and, in general, by the same reasoning,

$$E\left[\binom{N}{i}\right] = E[\text{number of sets of size } i \text{ that occur}]$$
$$= E\left[\sum\sum\sum\sum_{j_1<j_2<\cdots<j_i} I_{j_1} I_{j_2} \cdots I_{j_i}\right]$$
$$= \sum\sum\sum\sum_{j_1<j_2<\cdots<j_i} P(A_{j_1} A_{j_2} \cdots A_{j_i}).$$

Hence, (1.3.8) is a statement of the well-known identity

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum\sum_{i<j} P(A_i A_j) + \sum\sum\sum_{i<j<k} P(A_i A_j A_k)$$
$$- \cdots + (-1)^{n+1} P(A_1 A_2 \cdots A_n).$$

Other useful identities can also be derived by this approach. For instance, suppose we want a formula for the probability that exactly r of the events $A_1, \ldots, A_n$ occur. Then define

$$I_r = \begin{cases} 1 & \text{if } N = r \\ 0 & \text{otherwise} \end{cases}$$

and use the identity

$$\binom{N}{r} (1-1)^{N-r} = I_r$$

or

$$I_r = \binom{N}{r} \sum_{i=0}^{N-r} \binom{N-r}{i} (-1)^i$$
$$= \sum_{i=0}^{n-r} \binom{N}{r}\binom{N-r}{i} (-1)^i$$
$$= \sum_{i=0}^{n-r} \binom{N}{r+i}\binom{r+i}{r} (-1)^i.$$

Taking expectations of both sides of the above yields

$$E[I_r] = \sum_{i=0}^{n-r} (-1)^i \binom{r+i}{r} E\left[\binom{N}{r+i}\right]$$

or

$$(1.3.9)\quad P\{\text{exactly } r \text{ of the events } A_1, \ldots, A_n \text{ occur}\} = \sum_{i=0}^{n-r} (-1)^i \binom{r+i}{r} \sum_{j_1<j_2<\cdots<j_{r+i}} \cdots \sum P(A_{j_1}A_{j_2}\cdots A_{j_{r+i}}).$$

As an application of (1.3.9) suppose that m balls are randomly put in n boxes in such a way that, independent of the locations of the other balls, each ball is equally likely to go into any of the n boxes. Let us compute the probability that exactly r of the boxes are empty. By letting A_i denote the event that the ith box is empty, we see from (1.3.9) that

$$P\{\text{exactly } r \text{ of the boxes are empty}\} = \sum_{i=0}^{n-r} (-1)^i \binom{r+i}{r}\binom{n}{r+i}\left(1 - \frac{r+i}{n}\right)^m,$$

where the above follows since $\Sigma_{j_1<\cdots<j_{r+i}}$ consists of $\binom{n}{r+i}$ terms and each term in the sum is equal to the probability that a given set of $r + i$ boxes is empty.

Our next example illustrates what has been called the *probabilistic method*. This method, much employed and popularized by the mathematician Paul Erdos, attempts to solve deterministic problems by first introducing a probability structure and then employing probabilistic reasoning.

EXAMPLE 1.3(c) A graph is a set of elements, called nodes, and a set of (unordered) pairs of nodes, called edges. For instance, Figure 1.3.1 illustrates a graph with the set of nodes $N = \{1, 2, 3, 4, 5\}$ and the set of edges $E = \{(1, 2), (1, 3), (1, 5), (2, 3), (2, 4), (3, 4), (3, 5)\}$. Show that for any graph there is a subset of nodes A such that at least one-half of the edges have one of their nodes in A and the other in A^c. (For instance, in the graph illustrated in Figure 1.3.1 we could take $A = \{1, 2, 4\}$.)

Solution. Suppose that the graph contains m edges, and arbitrarily number them as $1, 2, \ldots, m$. For any set of nodes B, if we let $C(B)$ denote the number of edges that have exactly one of their nodes in B, then the problem is to show that $\max_B C(B) \geq m/2$. To verify this, let us introduce probability by randomly choosing a set of nodes S so that each node of the graph is independently in S with probability 1/2. If we now let X denote the number of edges in

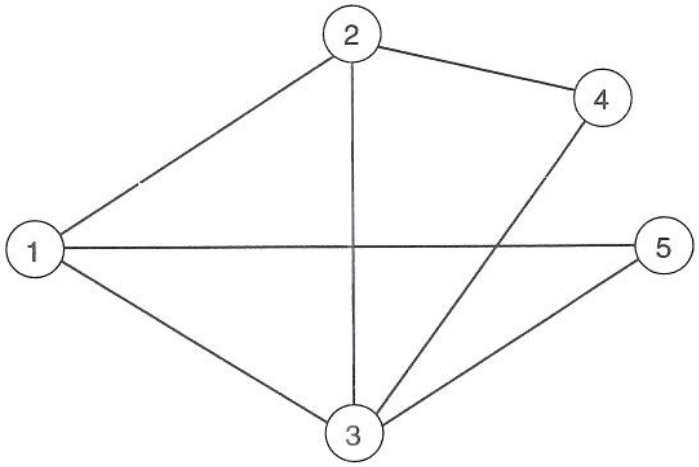

Figure 1.3.1. A graph.

the graph that have exactly one of their nodes in S, then X is a random variable whose set of possible values is all of the possible values of $C(B)$. Now, letting X_i equal 1 if edge i has exactly one of its nodes in S and letting it be 0 otherwise, then

$$E[X] = E\left[\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} E[X_i] = m/2.$$

Since at least one of the possible values of a random variable must be at least as large as its mean, we can thus conclude that $C(B) \geq m/2$ for some set of nodes B. (In fact, provided that the graph is such that $C(B)$ is not constant, we can conclude that $C(B) > m/2$ for some set of nodes B.)

Problems 1.9 and 1.10 give further applications of the probabilistic method.

1.4 Moment Generating, Characteristic Functions, and Laplace Transforms

The *moment generating function* of X is defined by

$$\psi(t) = E[e^{tX}]$$
$$= \int e^{tX}\, dF(x).$$

All the moments of X can be successively obtained by differentiating ψ and then evaluating at $t = 0$. That is,

$$\psi'(t) = E[Xe^{tX}].$$
$$\psi''(t) = E[X^2e^{tX}]$$
$$\vdots$$
$$\psi^n(t) = E[X^ne^{tX}].$$

Evaluating at $t = 0$ yields

$$\psi^n(0) = E[X^n], \qquad n \geq 1.$$

It should be noted that we have assumed that it is justifiable to interchange the differentiation and integration operations. This is usually the case.

When a moment generating function exists, it uniquely determines the distribution. This is quite important because it enables us to characterize the probability distribution of a random variable by its generating function.

EXAMPLE 1.4(A) Let X and Y be independent normal random variables with respective means μ_1 and μ_2 and respective variances

Table 1.4.1

Discrete Probability Distribution	Probability Mass Function, $p(x)$	Moment Generating Function, $\psi(t)$	Mean	Variance
Binomial with parameters n, p, $0 \leq p \leq 1$	$\binom{n}{x}p^x(1-p)^{n-x}$, $x = 0, 1, \ldots, n$	$(pe^t + (1-p))^n$	np	$np(1-p)$
Poisson with parameter $\lambda > 0$	$e^{-\lambda}\frac{\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$	$\exp\{\lambda(e^t - 1)\}$	λ	λ
Geometric with parameter $0 \leq p \leq 1$	$p(1-p)^{x-1}$, $x = 1, 2, \ldots$	$\frac{pe^t}{1-(1-p)e^t}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative binomial with parameters r, p	$\binom{x-1}{r-1}p^r(1-p)^{x-r}$, $x = r, r+1, \ldots$	$\left(\frac{pe^t}{1-(1-p)e^t}\right)^r$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$

σ_1^2 and σ_2^2. The moment generating function of their sum is given by

$$\begin{aligned}\psi_{X+Y}(t) &= E[e^{t(X+Y)}] \\ &= E[e^{tX}]E[e^{tY}] \qquad \text{(by independence)} \\ &= \psi_X(t)\psi_Y(t) \\ &= \exp\{(\mu_1 + \mu_2)t + (\sigma_1^2 + \sigma_2^2)t^2/2\},\end{aligned}$$

where the last equality comes from Table 1.4.2. Thus the moment generating function of $X + Y$ is that of a normal random variable with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. By uniqueness, this is the distribution of $X + Y$.

As the moment generating function of a random variable X need not exist, it is theoretically convenient to define the *characteristic function* of X by

$$\phi(t) = E[e^{itX}], \qquad -\infty < t < \infty,$$

where $i = \sqrt{-1}$. It can be shown that ϕ always exists and, like the moment generating function, uniquely determines the distribution of X.

We may also define the joint moment generating of the random variables $X_1, \ldots, X_n$ by

$$\psi(t_1, \ldots, t_n) = E\left[\exp\left\{\sum_{j=1}^{n} t_j X_j\right\}\right],$$

or the joint characteristic function by

$$\phi(t_1, \ldots, t_n) = E\left[\exp\left\{i\sum_{j=1}^{n} t_j X_j\right\}\right].$$

It may be proven that the joint moment generating function (when it exists) or the joint characteristic function uniquely determines the joint distribution.

Example 1.4(b) The Multivariate Normal Distribution. Let $Z_1, \ldots, Z_n$ be independent standard normal random variables. If for some constants a_{ij}, $1 \le i \le m$, $1 \le j \le n$, and μ_i, $1 \le i \le m$,

$$\begin{aligned}X_1 &= a_{11}Z_1 + \cdots + a_{1n}Z_n + \mu_1, \\ X_2 &= a_{21}Z_1 + \cdots + a_{2n}Z_n + \mu_2, \\ &\vdots \\ X_i &= a_{i1}Z_1 + \cdots + a_{in}Z_n + \mu_i, \\ &\vdots \\ X_m &= a_{m1}Z_1 + \cdots + a_{mn}Z_n + \mu_m\end{aligned}$$

Table 1.4.2

Continuous Probability Distribution	Probability Density Function, $f(x)$	Moment Generating Function, $\phi(t)$	Mean	Variance
Uniform over (a, b)	$\frac{1}{b-a}, a < x < b$	$\frac{e^{tb} - e^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential with parameter $\lambda > 0$	$\lambda e^{-\lambda x}, x \geq 0$	$\frac{\lambda}{\lambda - t}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma with parameters (n, λ), $\lambda > 0$	$\frac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!}, x \geq 0$	$\left(\frac{\lambda}{\lambda - t}\right)^n$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$
Normal with parameters (μ, σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < \infty$	$\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$	μ	σ^2
Beta with parameters a, b, $a > 0, b > 0$	$cx^{a-1}(1-x)^{b-1}, 0 < x < 1$ $c = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$		$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$

then the random variables $X_1, \ldots, X_m$ are said to have a multivariate normal distribution.

Let us now consider

$$\psi(t_1, \ldots, t_m) = E[\exp\{t_1 X_1 + \cdots + t_m X_m\}],$$

the joint moment generating function of $X_1, \ldots, X_m$. The first thing to note is that since $\sum_{i=1}^{m} t_i X_i$ is itself a linear combination of the independent normal random variables $Z_1, \ldots, Z_n$ it is also normally distributed. Its mean and variance are

$$E\left[\sum_{i=1}^{m} t_i X_i\right] = \sum_{i=1}^{m} t_i \mu_i$$

and

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{m} t_i X_i\right) &= \operatorname{Cov}\left(\sum_{i=1}^{m} t_i X_i, \sum_{j=1}^{m} t_j X_j\right) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} t_i t_j \operatorname{Cov}(X_i, X_j).\end{aligned}$$

Now, if Y is a normal random variable with mean μ and variance σ^2 then

$$E[e^Y] = \psi_Y(t)|_{t=1} = e^{\mu + \sigma^2/2}.$$

Thus, we see that

$$\psi(t_1, \ldots, t_m) = \exp\left\{\sum_{i=1}^{m} t_i \mu_i + 1/2 \sum_{i=1}^{m} \sum_{j=1}^{m} t_i t_j \operatorname{Cov}(X_i, X_j)\right\},$$

which shows that the joint distribution of $X_1, \ldots, X_m$ is completely determined from a knowledge of the values of $E[X_i]$ and $\operatorname{Cov}(X_i, X_j)$, $i, j = 1, \ldots, m$.

When dealing with random variables that only assume nonnegative values, it is sometimes more convenient to use *Laplace transforms* rather than characteristic functions. The Laplace transform of the distribution F is defined by

$$\tilde{F}(s) = \int_0^\infty e^{-sx}\, dF(x).$$

This integral exists for complex variables $s = a + bi$, where $a \geq 0$. As in the case of characteristic functions, the Laplace transform uniquely determines the distribution.

We may also define Laplace transforms for arbitrary functions in the following manner: The Laplace transform of the function g, denoted $\tilde{g}$, is defined by

$$\tilde{g}(s) = \int_0^\infty e^{-sx}\, dg(x)$$

provided the integral exists. It can be shown that $\tilde{g}$ determines g up to an additive constant.

1.5 Conditional Expectation

If X and Y are discrete random variables, the conditional probability mass function of X, given $Y = y$, is defined, for all y such that $P\{Y = y\} > 0$, by

$$P\{X = x \mid Y = y\} = \frac{P\{X = x, Y = y\}}{P\{Y = y\}}.$$

The conditional distribution function of X given $Y = y$ is defined by

$$F(x \mid y) = P\{X \leq x \mid Y = y\}$$

and the conditional expectation of X given $Y = y$, by

$$E[X \mid Y = y] = \int x\, dF(x \mid y) = \sum_x xP\{X = x \mid Y = y\}.$$

If X and Y have a joint probability density function $f(x, y)$, the conditional probability density function of X, given $Y = y$, is defined for all y such that $f_Y(y) > 0$ by

$$f(x \mid y) = \frac{f(x, y)}{f_Y(y)},$$

and the conditional probability distribution function of X, given $Y = y$, by

$$F(x \mid y) = P\{X \leq x \mid Y = y\} = \int_{-\infty}^{x} f(x \mid y)\, dx.$$

The conditional expectation of X, given $Y = y$, is defined, in this case, by

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} xf(x \mid y)\, dx.$$

Thus all definitions are exactly as in the unconditional case except that all probabilities are now conditional on the event that $Y = y$.

Let us denote by $E[X|Y]$ that function of the random variable Y whose value at $Y = y$ is $E[X|Y = y]$. An extremely useful property of conditional expectation is that for all random variables X and Y

$$E[X] = E[E[X|Y]] = \int E[X|Y = y]\, dF_Y(y) \tag{1.5.1}$$

when the expectations exist.

If Y is a discrete random variable, then Equation (1.5.1) states

$$E[X] = \sum_y E[X|Y = y]P\{Y = y\},$$

While if Y is continuous with density $f(y)$, then Equation (1.5.1) says

$$E[X] = \int_{-\infty}^{\infty} E[X|Y = y]f(y)\, dy.$$

We now give a proof of Equation (1.5.1) in the case where X and Y are both discrete random variables.

Proof of (1.5.1) *when X and Y Are Discrete* To show

$$E[X] = \sum_y E[X|Y = y]P\{Y = y\}.$$

We write the right-hand side of the above as

$$\begin{aligned}
\sum_y E[X|Y = y]P\{Y = y\} &= \sum_y \sum_x xP\{X = x|Y = y\}P\{Y = y\} \\
&= \sum_y \sum_x xP\{X = x, Y = y\} \\
&= \sum_x x \sum_y P\{X = x, Y = y\} \\
&= \sum_x xP\{X = x\} \\
&= E[X].
\end{aligned}$$

and the result is obtained.

Thus from Equation (1.5.1) we see that $E[X]$ is a weighted average of the conditional expected value of X given that $Y = y$, each of the terms $E[X|Y = y]$ being weighted by the probability of the event on which it is conditioned.

EXAMPLE 1.5(A) ***The Sum of a Random Number of Random Variables.*** Let $X_1, X_2, \ldots$ denote a sequence of independent and identically distributed random variables; and let N denote a nonnegative integer valued random variable that is independent of the sequence $X_1, X_2, \ldots$. We shall compute the moment generating function of $Y = \sum_1^N X_i$ by first conditioning on N. Now

$$\begin{aligned} &E\left[\exp\left\{t\sum_1^N X_i\right\}\middle| N = n\right] \\ &\quad = E\left[\exp\left\{t\sum_1^n X_i\right\}\middle| N = n\right] \\ &\quad = E\left[\exp\left\{t\sum_1^n X_i\right\}\right] \qquad \text{(by independence)} \\ &\quad = (\Psi_X(t))^n, \end{aligned}$$

where $\Psi_X(t) = E[e^{tX}]$ is the moment generating function of X. Hence,

$$E\left[\exp\left\{t\sum_1^N X_i\right\}\middle| N\right] = (\psi_X(t))^N$$

and so

$$\psi_Y(t) = E\left[\exp\left\{t\sum_1^N X_i\right\}\right] = E[(\psi_X(t))^N].$$

To compute the mean and variance of $Y = \sum_1^N X_i$, we differentiate $\psi_Y(t)$ as follows:

$$\begin{aligned} \psi_Y'(t) &= E[N(\psi_X(t))^{N-1}\psi_X'(t)], \\ \psi_Y''(t) &= E[N(N-1)(\psi_X(t))^{N-2}(\psi_X'(t))^2 + N(\psi_X(t))^{N-1}\psi_X''(t)]. \end{aligned}$$

Evaluating at $t = 0$ gives

$$E[Y] = E[NE[X]] = E[N]E[X]$$

and

$$\begin{aligned} E[Y^2] &= E[N(N-1)E^2[X] + NE[X^2]] \\ &= E[N]\,\mathrm{Var}(X) + E[N^2]E^2[X]. \end{aligned}$$

N
priat
biliti

Hence,

$$\begin{aligned}\mathrm{Var}(Y) &= E[Y^2] - E^2[Y] \\ &= E[N]\,\mathrm{Var}(X) + E^2[X]\,\mathrm{Var}(N).\end{aligned}$$

EXAMPLE 1.5(B) A miner is trapped in a mine containing three doors. The first door leads to a tunnel that takes him to safety after two hours of travel. The second door leads to a tunnel that returns him to the mine after three hours of travel. The third door leads to a tunnel that returns him to his mine after five hours. Assuming that the miner is at all times equally likely to choose any one of the doors, let us compute the moment generating function of X, the time when the miner reaches safety.

Let Y denote the door initially chosen. Then

$$E[e^{tX}] = \tfrac{1}{3}(E[e^{tX} \mid Y = 1] + E[e^{tX} \mid Y = 2] + E[e^{tX} \mid Y = 3]). \tag{1.5.2}$$

Now given that $Y = 1$, it follows that $X = 2$, and so

$$E[e^{tX} \mid Y = 1] = e^{2t}.$$

Also, given that $Y = 2$, it follows that $X = 3 + X'$, where X' is the number of additional hours to safety after returning to the mine. But once the miner returns to his cell the problem is exactly as before, and thus X' has the same distribution as X. Therefore,

$$\begin{aligned}E[e^{tX} \mid Y = 2] &= E[e^{t(3+X')}] \\ &= e^{3t}E[e^{tX}].\end{aligned}$$

Similarly,

$$E[e^{tX} \mid Y = 3] = e^{5t}E[e^{tX}].$$

Substitution back into (1.5.2) yields

$$E[e^{tX}] = \tfrac{1}{3}(e^{2t} + e^{3t}E[e^{tX}] + e^{5t}E[e^{tX}])$$

or

$$E[e^{tX}] = \frac{e^{2t}}{3 - e^{3t} - e^{5t}}.$$

Not only can we obtain expectations by first conditioning upon an appropriate random variable, but we may also use this approach to compute probabilities. To see this, let E denote an arbitrary event and define the indicator

random variable X by

$$X = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{if } E \text{ does not occur.} \end{cases}$$

It follows from the definition of X that

$$E[X] = P(E)$$
$$E[X|Y = y] = P(E|Y = y) \qquad \text{for any random variable } Y.$$

Therefore, from Equation (1.5.1) we obtain that

$$P(E) = \int P(E|Y = y)\, dF_Y(y).$$

Example 1.5(c) Suppose in the matching problem, Example 1.3(a), that those choosing their own hats depart, while the others (those without a match) put their selected hats in the center of the room, mix them up, and then reselect. If this process continues until each individual has his or her own hat, find $E[R_n]$ where R_n is the number of rounds that are necessary.

We will now show that $E[R_n] = n$. The proof will be by induction on n, the number of individuals. As it is obvious for $n = 1$ assume that $E[R_k] = k$ for $k = 1, \ldots, n-1$. To compute $E[R_n]$, start by conditioning on M, the number of matches that occur in the first round. This gives

$$E[R_n] = \sum_{i=0}^{n} E[R_n|M = i]P\{M = i\}.$$

Now, given a total of i matches in the initial round, the number of rounds needed will equal 1 plus the number of rounds that are required when $n - i$ people remain to be matched with their hats. Therefore,

$$\begin{aligned} E[R_n] &= \sum_{i=0}^{n} (1 + E[R_{n-i}])P\{M = i\} \\ &= 1 + E[R_n]P\{M = 0\} + \sum_{i=1}^{n} E[R_{n-i}]P\{M = i\} \\ &= 1 + E[R_n]P\{M = 0\} + \sum_{i=1}^{n} (n - i)P\{M = i\} \\ &\qquad \text{(by the induction hypothesis)} \\ &= 1 + E[R_n]P\{M = 0\} + n(1 - P\{M = 0\}) - E[M] \\ &= E[R_n]P\{M = 0\} + n(1 - P\{M = 0\}) \\ &\qquad \text{(since } E[M] = 1\text{)} \end{aligned}$$

which proves the result.

EXAMPLE 1.5(D) Suppose that X and Y are independent random variables having respective distributions F and G. Then the distribution of $X + Y$—which we denote by $F * G$, and call the *convolution* of F and G—is given by

$$\begin{aligned}(F * G)(a) &= P\{X + Y \le a\} \\ &= \int_{-\infty}^{\infty} P\{X + Y \le a \mid Y = y\}\, dG(y) \\ &= \int_{-\infty}^{\infty} P\{X + y \le a \mid Y = y\}\, dG(y) \\ &= \int_{-\infty}^{\infty} F(a - y)\, dG(y).\end{aligned}$$

We denote $F * F$ by F_2 and in general $F * F_{n-1} = F_n$. Thus F_n, the n-fold convolution of F with itself, is the distribution of the sum of n independent random variables each having distribution F.

EXAMPLE 1.5(E) ***The Ballot Problem.*** In an election, candidate A receives n votes and candidate B receives m votes, where $n > m$. Assuming that all orderings are equally likely, show that the probability that A is always ahead in the count of votes is $(n - m)/(n + m)$.

Solution. Let $P_{n,m}$ denote the desired probability. By conditioning on which candidate receives the last vote counted we have

$$\begin{aligned}P_{n,m} &= P\{A \text{ always ahead} \mid A \text{ receives last vote}\}\frac{n}{n+m} \\ &\quad + P\{A \text{ always ahead} \mid B \text{ receives last vote}\}\frac{m}{n+m}.\end{aligned}$$

Now it is easy to see that, given that A receives the last vote, the probability that A is always ahead is the same as if A had received a total of $n - 1$ and B a total of m votes. As a similar result is true when we are given that B receives the last vote, we see from the above that

$$P_{n,m} = \frac{n}{n+m}P_{n-1,m} + \frac{m}{m+n}P_{n,m-1}. \tag{1.5.3}$$

We can now prove that

$$P_{n,m} = \frac{n-m}{n+m}$$

by induction on $n + m$. As it is obviously true when $n + m = 1$—that is, $P_{1,0} = 1$—assume it whenever $n + m = k$. Then when

$n + m = k + 1$ we have by (1.5.3) and the induction hypothesis

$$P_{n,m} = \frac{n}{n+m}\frac{n-1-m}{n-1+m} + \frac{m}{m+n}\frac{n-m+1}{n+m-1}$$
$$= \frac{n-m}{n+m}.$$

The result is thus proven.

The ballot problem has some interesting applications. For example, consider successive flips of a coin that always lands on "heads" with probability p, and let us determine the probability distribution of the first time, after beginning, that the total number of heads is equal to the total number of tails. The probability that the first time this occurs is at time $2n$ can be obtained by first conditioning on the total number of heads in the first $2n$ trials. This yields

$$P\{\text{first time equal} = 2n\}$$
$$= P\{\text{first time equal} = 2n \mid n \text{ heads in first } 2n\}\binom{2n}{n}p^n(1-p)^n.$$

Now given a total of n heads in the first $2n$ flips, it is easy to see that all possible orderings of the n heads and n tails are equally likely and thus the above conditional probability is equivalent to the probability that in an election in which each candidate receives n votes, one of the candidates is always ahead in the counting until the last vote (which ties them). But by conditioning on whoever receives the last vote, we see that this is just the probability in the ballot problem when $m = n - 1$. Hence,

$$P\{\text{first time equal} = 2n\} = P_{n,n-1}\binom{2n}{n}p^n(1-p)^n$$
$$= \frac{\binom{2n}{n}p^n(1-p)^n}{2n-1}.$$

EXAMPLE 1.5(F) ***The Matching Problem Revisited.*** Let us reconsider Example 1.3(a) in which n individuals mix their hats up and then randomly make a selection. We shall compute the probability of exactly k matches.

First let E denote the event that no matches occur, and to make explicit the dependence on n write $P_n = P(E)$. Upon conditioning on whether or not the first individual selects his or her own hat—call these events M and M^c—we obtain

$$P_n = P(E) = P(E|M)P(M) + P(E|M^c)P(M^c).$$

Clearly, $P(E|M) = 0$, and so

$$P_n = P(E|M^c)\frac{n-1}{n}. \tag{1.5.4}$$

Now, $P(E|M^c)$ is the probability of no matches when $n-1$ people select from a set of $n-1$ hats that does not contain the hat of one of them. This can happen in either of two mutually exclusive ways. Either there are no matches and the extra person does not select the extra hat (this being the hat of the person that chose first), or there are no matches and the extra person does select the extra hat. The probability of the first of these events is P_{n-1}, which is seen by regarding the extra hat as "belonging" to the extra person. Since the second event has probability $[1/(n-1)]P_{n-2}$, we have

$$P(E|M^c) = P_{n-1} + \frac{1}{n-1}P_{n-2}$$

and thus, from Equation (1.5.4),

$$P_n = \frac{n-1}{n}P_{n-1} + \frac{1}{n}P_{n-2},$$

or, equivalently,

$$P_n - P_{n-1} = -\frac{1}{n}(P_{n-1} - P_{n-2}). \tag{1.5.5}$$

However, clearly

$$P_1 = 0, \qquad P_2 = \tfrac{1}{2}.$$

Thus, from Equation (1.5.5),

$$P_3 - P_2 = -\frac{(P_2 - P_1)}{3} = -\frac{1}{3!} \quad \text{or} \quad P_3 = \frac{1}{2!} - \frac{1}{3!},$$

$$P_4 - P_3 = -\frac{(P_3 - P_2)}{4} = \frac{1}{4!} \quad \text{or} \quad P_4 = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!},$$

and, in general, we see that

$$P_n = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots + \frac{(-1)^n}{n!}.$$

To obtain the probability of exactly k matches, we consider any fixed group of k individuals. The probability that they, and only

they, select their own hats is

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

where P_{n-k} is the conditional probability that the other $n-k$ individuals, selecting among their own hats, have no matches. As there are $\binom{n}{k}$ choices of a set of k individuals, the desired probability of exactly k matches is

$$\binom{n}{k}\frac{(n-k)!}{n!}P_{n-k}=\frac{\frac{1}{2!}-\frac{1}{3!}+\cdots+\frac{(-1)^{n-k}}{(n-k)!}}{k!},$$

which, for n large, is approximately equal to $e^{-1}/k!$.

Thus for n large the number of matches has approximately the Poisson distribution with mean 1. To understand this result better recall that the Poisson distribution with mean λ is the limiting distribution of the number of successes in n independent trials, each resulting in a success with probability p_n, when $np_n \to \lambda$ as $n \to \infty$. Now if we let

$$X_i=\begin{cases}1 & \text{if the } i\text{th person selects his or her own hat}\\ 0 & \text{otherwise,}\end{cases}$$

then the number of matches, $\sum_{i=1}^{n} X_i$, can be regarded as the number of successes in n trials when each is a success with probability $1/n$. Now, whereas the above result is not immediately applicable because these trials are not independent, it is true that it is a rather weak dependence since, for example,

$$P\{X_i=1\}=1/n$$

and

$$P\{X_i=1\,|\,X_j=1\}=1/(n-1), \qquad j\neq i.$$

Hence we would certainly hope that the Poisson limit would still remain valid under this type of weak dependence. The results of this example show that it does.

EXAMPLE 1.5(G) ***A Packing Problem.*** Suppose that n points are arranged in linear order, and suppose that a pair of adjacent points is chosen at random. That is, the pair $(i, i+1)$ is chosen with probability $1/(n-1)$, $i=1, 2, \ldots, n-1$. We then continue to randomly choose pairs, disregarding any pair having a point

Figure 1.5.1

previously chosen, until only isolated points remain. We are interested in the mean number of isolated points.

For instance, if $n = 8$ and the random pairs are, in order of appearance, (2, 3), (7, 8), (3, 4), and (4, 5), then there will be two isolated points (the pair (3, 4) is disregarded) as shown in Figure 1.5.1.

If we let

$$I_{i,n} = \begin{cases} 1 & \text{if point } i \text{ is isolated} \\ 0 & \text{otherwise,} \end{cases}$$

then $\sum_{i=1}^{n} I_{i,n}$ represents the number of isolated points. Hence

$$E[\text{number of isolated points}] = \sum_{i=1}^{n} E[I_{i,n}]$$

$$= \sum_{i=1}^{n} P_{i,n},$$

where $P_{i,n}$ is defined to be the probability that point i is isolated when there are n points. Let

$$P_n \equiv P_{n,n} = P_{1,n}.$$

That is, P_n is the probability that the extreme point n (or 1) will be isolated. To derive an expression for $P_{i,n}$, note that we can consider the n points as consisting of two contiguous segments, namely,

$$1, 2, \ldots, i \quad \text{and} \quad i, i+1, \ldots, n.$$

Since point i will be vacant if and only if the right-hand point of the first segment and the left-hand point of the second segment are both vacant, we see that

$$P_{i,n} = P_i P_{n-i+1}. \tag{1.5.6}$$

Hence the $P_{i,n}$ will be determined if we can calculate the corresponding probabilities that extreme points will be vacant. To derive an expression for the P_n condition on the initial pair—say $(i, i+1)$—and note that this choice breaks the line into two independent segments—$1, 2, \ldots, i-1$ and $i+2, \ldots, n$. That is, if the initial

pair is $(i, i+1)$, then the extreme point n will be isolated if the extreme point of a set of $n-i-1$ points is isolated. Hence we have

$$P_n = \sum_{i=1}^{n-1} \frac{P_{n-i-1}}{n-1} = \frac{P_1 + \cdots + P_{n-2}}{n-1}$$

or

$$(n-1)P_n = P_1 + \cdots + P_{n-2}.$$

Substituting $n-1$ for n gives

$$(n-2)P_{n-1} = P_1 + \cdots + P_{n-3}$$

and subtracting these two equations gives

$$(n-1)P_n - (n-2)P_{n-1} = P_{n-2}$$

or

$$P_n - P_{n-1} = -\frac{P_{n-1} - P_{n-2}}{n-1}.$$

Since $P_1 = 1$ and $P_2 = 0$, this yields

$$P_3 - P_2 = -\frac{P_2 - P_1}{2} = \frac{1}{2} \quad \text{or} \quad P_3 = \frac{1}{2!}$$

$$P_4 - P_3 = -\frac{P_3 - P_2}{3} = -\frac{1}{3!} \quad \text{or} \quad P_4 = \frac{1}{2!} - \frac{1}{3!},$$

and, in general,

$$P_n = \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^{n-1}}{(n-1)!} = \sum_{j=0}^{n-1} \frac{(-1)^j}{j!}, \qquad n \geq 2.$$

Thus, from (1.5.6),

$$P_{i,n} = \begin{cases} \displaystyle\sum_{j=0}^{n-1} \frac{(-1)^j}{j!} & i = 1, n \\ 0 & i = 2, n-1 \\ \displaystyle\sum_{j=0}^{i-1} \frac{(-1)^j}{j!} \sum_{j=0}^{n-i} \frac{(-1)^j}{j!} & 2 < i < n-1. \end{cases}$$

For i and $n - i$ large we see from the above that $P_{i,n} \approx e^{-2}$, and, in fact, it can be shown from the above that $\sum_{i=1}^{n} P_{i,n}$—the expected number of vacant points—is approximately given by

$$\sum_{i=1}^{n} P_{i,n} \approx (n+2)e^{-2} \qquad \text{for large } n.$$

EXAMPLE 1.5(H) ***A Reliability Example.*** Consider an n component system that is subject to randomly occurring shocks. Suppose that each shock has a value that, independent of all else, is chosen from a distribution G. If a shock of value x occurs, then each component that was working at the moment the shock arrived will, independently, instantaneously fail with probability x. We are interested in the distribution of N, the number of necessary shocks until all components are failed.

To compute $P\{N > k\}$ let E_i, $i = 1, \ldots, n$, denote the event that component i has survived the first k shocks. Then

$$\begin{aligned} P\{N > k\} &= P\left(\bigcup_{1}^{n} E_i\right) \\ &= \sum_i P(E_i) - \sum_{i<l} P(E_iE_l) \\ &\quad + \cdots + (-1)^{n+1} P(E_1E_2 \cdots E_n). \end{aligned}$$

To compute the above probability let p_j denote the probability that a given set of j components will all survive some arbitrary shock. Conditioning on the shock's value gives

$$\begin{aligned} p_j &= \int P\{\text{set of } j \text{ survive} \mid \text{value is } x\}\, dG(x) \\ &= \int (1-x)^j\, dG(x). \end{aligned}$$

Since

$$\begin{aligned} P(E_i) &= p_1^k, \\ P(E_iE_l) &= p_2^k, \ldots, P(E_1 \cdots E_n) = p_n^k, \end{aligned}$$

we see that

$$P\{N > k\} = np_1^k - \binom{n}{2} p_2^k + \binom{n}{3} p_3^k \cdots (-1)^{n+1} p_n^k.$$

The mean of N can be computed from the above as follows:

$$\begin{aligned} E[N] &= \sum_{k=0}^{\infty} P\{N > k\} \\ &= \sum_{k=0}^{\infty} \sum_{i=1}^{n} \binom{n}{i} (-1)^{i+1} p_i^k \\ &= \sum_{i=1}^{n} \binom{n}{i} (-1)^{i+1} \sum_{k=0}^{\infty} p_i^k \\ &= \sum_{i=1}^{n} \binom{n}{i} \frac{(-1)^{i+1}}{1 - p_i}. \end{aligned}$$

The reader should note that we have made use of the identity $E[N] = \sum_{k=0}^{\infty} P\{N > k\}$, valid for all nonnegative integer-valued random variables N (see Problem 1.1).

EXAMPLE 1.5(I) ***Classifying a Poisson Number of Events.*** Suppose that we are observing events, and that N, the total number that occur, is a Poisson random variable with mean λ. Suppose also that each event that occurs is, independent of other events, classified as a type j event with probability p_j, $j = 1, \ldots, k$, $\sum_{j=1}^{k} p_j = 1$. Let N_j denote the number of type j events that occur, $j = 1, \ldots, k$, and let us determine their joint probability mass function.

For any nonnegative integers n_j, $j = 1, \ldots, k$, let $n = \sum_{j=1}^{k} n_j$. Then, since $N = \sum_j N_j$, we have that

$$\begin{aligned} &P\{N_j = n_j, j = 1, \ldots, k\} \\ &\quad = P\{N_j = n_j, j = 1, \ldots, k \mid N = n\} P\{N = n\} \\ &\qquad + P\{N_j = n_j, j = 1, \ldots, k \mid N \neq n\} P\{N \neq n\} \\ &\quad = P\{N_j = n_j, j = 1, \ldots, k \mid N = n\} P\{N = n\}. \end{aligned}$$

Now, given that there are a total of $N = n$ events it follows, since each event is independently a type j event with probability p_j, $1 \leq j \leq k$, that $N_1, N_2, \ldots, N_k$ has a multinomial distribution with parameters n and $p_1, p_2 \ldots, p_k$. Therefore,

$$\begin{aligned} P\{N_j = n_j, j = 1, \ldots, k\} &= \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} e^{-\lambda} \lambda^n / n! \\ &= \prod_j e^{-\lambda p_j} \frac{(\lambda p_j)^{n_j}}{n_j!}. \end{aligned}$$

Thus we can conclude that the N_j are independent Poisson random variables with respective means λp_j, $j = 1, \ldots, k$.

Conditional expectations given that $Y = y$ satisfy all of the properties of ordinary expectations, except that now all probabilities are conditioned on the event that $\{Y = y\}$. Hence, we have that

$$E\left[\sum_{i=1}^{n} X_i \mid Y = y\right] = \sum_{i=1}^{n} E[X_i \mid Y = y]$$

implying that

$$E\left[\sum_{i=1}^{n} X_i \mid Y\right] = \sum_{i=1}^{n} E[X_i \mid Y].$$

Also, from the equality $E[X] = E[E[X \mid Y]]$ we can conclude that

$$E[X \mid W = w] = E[E[X \mid W = w, Y] \mid W = w]$$

or, equivalently,

$$E[X \mid W] = E[E[X \mid W, Y] \mid W].$$

Also, we should note that the fundamental result

$$E[X] = E[E[X \mid Y]]$$

remains valid even when Y is a random vector.

1.5.1 Conditional Expectations and Bayes Estimators

Conditional expectations have important uses in the Bayesian theory of statistics. A classical problem in this area arises when one is to observe data $\boldsymbol{X} = (X_1, \ldots, X_n)$ whose distribution is determined by the value of a random variable θ, which has a specified probability distribution (called the prior distribution). Based on the value of the data $\boldsymbol{X}$ a problem of interest is to estimate the unseen value of θ. An estimator of θ can be any function $d(\boldsymbol{X})$ of the data, and in Bayesian statistics one often wants to choose $d(\boldsymbol{X})$ to minimize $E[(d(\boldsymbol{X}) - \theta)^2 \mid \boldsymbol{X}]$, the conditional expected squared distance between the estimator and the parameter. Using the facts that

(i) conditional on $\boldsymbol{X}$, $d(\boldsymbol{X})$ is a constant; and

(ii) for any random variable W, $E[(W - c)^2]$ is minimized when $c = E[W]$

it follows that the estimator that minimizes $E[(d(\boldsymbol{X}) - \theta)^2 | \boldsymbol{X}]$, called the *Bayes estimator*, is given by

$$d(\boldsymbol{X}) = E[\theta | \boldsymbol{X}].$$

An estimator $d(\boldsymbol{X})$ is said to be an *unbiased* estimator of θ if

$$E[d(\boldsymbol{X}) | \theta] = \theta.$$

An important result in Bayesian statistics is that the only time that a Bayes estimator is unbiased is in the trivial case where it is equal to θ with probability 1. To prove this, we start with the following lemma.

Lemma 1.5.1

For any random variable Y and random vector $\boldsymbol{Z}$

$$E[(Y - E[Y|\boldsymbol{Z}])E[Y|\boldsymbol{Z}]] = 0.$$

Proof

$$\begin{aligned} E[YE[Y|\boldsymbol{Z}]] &= E[E[YE[Y|\boldsymbol{Z}]|\boldsymbol{Z}]] \\ &= E[E[Y|\boldsymbol{Z}]E[Y|\boldsymbol{Z}]] \end{aligned}$$

where the final equality follows because, given $\boldsymbol{Z}$, $E[Y|\boldsymbol{Z}]$ is a constant and so $E[YE[Y|\boldsymbol{Z}]|\boldsymbol{Z}] = E[Y|\boldsymbol{Z}]E[Y|\boldsymbol{Z}]$. Since the final equality is exactly what we wanted to prove, the lemma follows.

PROPOSITION 1.5.2

If $P\{E[\theta|\boldsymbol{X}] = \theta\} \neq 1$ then the Bayes estimator $E[\theta|\boldsymbol{X}]$ is not unbiased.

Proof Letting $Y = \theta$ and $\boldsymbol{Z} = \boldsymbol{X}$ in Lemma 1.5.1 yields that

$$E[(\theta - E[\theta|\boldsymbol{X}])E[\theta|\boldsymbol{X}]] = 0. \tag{1.5.7}$$

Now let $Y = E[\theta|\boldsymbol{X}]$ and suppose that Y is an unbiased estimator of θ so that $E[Y|\theta] = \theta$. Letting $\boldsymbol{Z} = \theta$ we obtain from Lemma 1.5.1 that

$$E[(E[\theta|\boldsymbol{X}] - \theta)\theta] = 0. \tag{1.5.8}$$

Upon adding Equations (1.5.7) and (1.5.8) we obtain that

$$E[(\theta - E[\theta|\boldsymbol{X}])E[\theta|\boldsymbol{X}]] + E[(E[\theta|\boldsymbol{X}] - \theta)\theta] = 0.$$

or,

$$E[(\theta - E[\theta|X])E[\theta|X] + (E[\theta|X] - \theta)\theta] = 0$$

or,

$$-E[(\theta - E[\theta|X])^2] = 0$$

implying that, with probability 1, $\theta - E[\theta|X] = 0$.

1.6 The Exponential Distribution, Lack of Memory, and Hazard Rate Functions

A continuous random variable X is said to have an *exponential distribution* with parameter λ, $\lambda > 0$, if its probability density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0, \end{cases}$$

or, equivalently, if its distribution is

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

The moment generating function of the exponential distribution is given by

$$E[e^{tX}] = \int_0^\infty e^{tx}\lambda e^{-\lambda x}\,dx = \frac{\lambda}{\lambda - t}. \tag{1.6.1}$$

All the moments of X can now be obtained by differentiating (1.6.1), and we leave it to the reader to verify that

$$E[X] = 1/\lambda, \qquad \mathrm{Var}(X) = 1/\lambda^2.$$

The usefulness of exponential random variables derives from the fact that they possess the memoryless property, where a random variable X is said to be without memory, or *memoryless*, if

$$P\{X > s + t | X > t\} = P\{X > s\} \qquad \text{for } s, t \geq 0. \tag{1.6.2}$$

If we think of X as being the lifetime of some instrument, then (1.6.2) states that the probability that the instrument lives for at least $s + t$ hours, given that it has survived t hours, is the same as the initial probability that it lives for at least s hours. In other words, if the instrument is alive at time t, then the distribution of its remaining life is the original lifetime distribution. The condition (1.6.2) is equivalent to

$$\bar{F}(s + t) = \bar{F}(s)\bar{F}(t),$$

and since this is satisfied when F is the exponential, we see that such random variables are memoryless.

Example 1.6(A) Consider a post office having two clerks, and suppose that when A enters the system he discovers that B is being served by one of the clerks and C by the other. Suppose also that A is told that his service will begin as soon as either B or C leaves. If the amount of time a clerk spends with a customer is exponentially distributed with mean $1/\lambda$, what is the probability that, of the three customers, A is the last to leave the post office?

The answer is obtained by reasoning as follows: Consider the time at which A first finds a free clerk. At this point either B or C would have just left and the other one would still be in service. However, by the lack of memory of the exponential, it follows that the amount of additional time that this other person has to spend in the post office is exponentially distributed with mean $1/\lambda$. That is, it is the same as if he was just starting his service at this point. Hence, by symmetry, the probability that he finishes before A must equal $\frac{1}{2}$.

Example 1.6(B) Let $X_1, X_2, \ldots$ be independent and identically distributed continuous random variables with distribution F. We say that a record occurs at time n, $n > 0$, and has value X_n if $X_n > \max(X_1, \ldots, X_{n-1})$, where $X_0 = -\infty$. That is, a record occurs each time a new high is reached. Let τ_i denote the time between the ith and the $(i + 1)$th record. What is its distribution?

As a preliminary to computing the distribution of τ_i, let us note that the record times of the sequence $X_1, X_2, \ldots$ will be the same as for the sequence $F(X_1), F(X_2), \ldots$, and since $F(X)$ has a uniform $(0, 1)$ distribution (see Problem 1.2), it follows that the distribution of τ_i does not depend on the actual distribution F (as long as it is continuous). So let us suppose that F is the exponential distribution with parameter $\lambda = 1$.

To compute the distribution of τ_i, we will condition on R_i the ith record value. Now $R_1 = X_1$ is exponential with rate 1. R_2 has the distribution of an exponential with rate 1 given that it is greater than R_1. But by the lack of memory property of the exponential this

means that R_2 has the same distribution as R_1 plus an independent exponential with rate 1. Hence R_2 has the same distribution as the sum of two independent exponential random variables with rate 1. The same argument shows that R_i has the same distribution as the sum of i independent exponentials with rate 1. But it is well known (see Problem 1.29) that such a random variable has the gamma distribution with parameters $(i, 1)$. That is, the density of R_i is given by

$$f_{R_i}(t) = \frac{e^{-t}t^{i-1}}{(i-1)!}, \qquad t \geq 0.$$

Hence, conditioning on R_i yields

$$\begin{aligned} P\{\tau_i > k\} &= \int_0^\infty P\{\tau_i > k \mid R_i = t\} \frac{e^{-t}t^{i-1}}{(i-1)!}\, dt \\ &= \int_0^\infty (1 - e^{-t})^k e^{-t} \frac{t^{i-1}}{(i-1)!}\, dt, \qquad i \geq 1, \end{aligned}$$

where the last equation follows since if the ith record value equals t, then none of the next k values will be records if they are all less than t.

It turns out that not only is the exponential distribution "memoryless," but it is the unique distribution possessing this property. To see this, suppose that X is memoryless and let $\bar{F}(x) = P\{X > x\}$. Then

$$\bar{F}(s + t) = \bar{F}(s)\bar{F}(t).$$

That is, $\bar{F}$ satisfies the functional equation

$$g(s + t) = g(s)g(t).$$

However, the only solutions of the above equation that satisfy any sort of reasonable condition (such as monotonicity, right or left continuity, or even measurability) are of the form

$$g(x) = e^{-\lambda x}$$

for some suitable value of λ. [A simple proof when g is assumed right continuous is as follows: Since $g(s + t) = g(s)g(t)$, it follows that $g(2/n) = g(1/n + 1/n) = g^2(1/n)$. Repeating this yields $g(m/n) = g^m(1/n)$. Also $g(1) = g(1/n + \cdots + 1/n) = g^n(1/n)$. Hence, $g(m/n) = (g(1))^{m/n}$, which implies, since g is right continuous, that $g(x) = (g(1))^x$. Since $g(1) = g^2(1/2) \geq 0$, we obtain $g(x) = e^{-\lambda x}$, where $\lambda = -\log(g(1))$]. Since a distribution function is always

right continuous, we must have

$$\bar{F}(x) = e^{-\lambda x}.$$

The memoryless property of the exponential is further illustrated by the failure rate function (also called the hazard rate function) of the exponential distribution.

Consider a continuous random variable X having distribution function F and density f. The *failure* (or *hazard*) *rate* function $\lambda(t)$ is defined by

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)}. \tag{1.6.3}$$

To interpret $\lambda(t)$, think of X as being the lifetime of some item, and suppose that X has survived for t hours and we desire the probability that it will not survive for an additional time dt. That is, consider $P\{X \in (t, t+dt) \mid X > t\}$. Now

$$\begin{aligned} P\{X \in (t, t+dt) \mid X > t\} &= \frac{P\{X \in (t, t+dt), X > t\}}{P\{X > t\}} \\ &= \frac{P\{X \in (t, t+dt)\}}{P\{X > t\}} \\ &\approx \frac{f(t)\, dt}{\bar{F}(t)} \\ &= \lambda(t)\, dt. \end{aligned}$$

That is, $\lambda(t)$ represents the probability intensity that a t-year-old item will fail.

Suppose now that the lifetime distribution is exponential. Then, by the memoryless property, it follows that the distribution of remaining life for a t-year-old item is the same as for a new item. Hence $\lambda(t)$ should be constant. This checks out since

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

Thus, the failure rate function for the exponential distribution is constant. The parameter λ is often referred to as the *rate* of the distribution. (Note that the rate is the reciprocal of the mean, and vice versa.)

It turns out that the failure rate function $\lambda(t)$ uniquely determines the distribution F. To prove this, we note that

$$\lambda(t) = \frac{-\frac{d}{dt}\bar{F}(t)}{\bar{F}(t)}.$$

Integration yields

$$\log \bar{F}(t) = -\int_0^t \lambda(t)\, dt + k$$

or

$$\bar{F}(t) = c \exp\left\{-\int_0^t \lambda(t)\, dt\right\}.$$

Letting $t = 0$ shows that $c = 1$ and so

$$\bar{F}(t) = \exp\left\{-\int_0^t \lambda(t)\, dt\right\}.$$

1.7 Some Probability Inequalities

We start with an inequality known as Markov's inequality.

Lemma 1.7.1 Markov's Inequality

If X is a nonnegative random variable, then for any $a > 0$

$$P\{X \geq a\} \leq E[X]/a.$$

Proof Let $I\{X \geq a\}$ be 1 if $X \geq a$ and 0 otherwise. Then, it is easy to see since $X \geq 0$ that

$$aI\{X \geq a\} \leq X.$$

Taking expectations yields the result.

PROPOSITION 1.7.2 Chernoff Bounds

Let X be a random variable with moment generating function $M(t) = E[e^{tX}]$. Then for $a > 0$

$$P\{X \geq a\} \leq e^{-ta}M(t) \qquad \text{for all } t > 0$$
$$P\{X \leq a\} \leq e^{-ta}M(t) \qquad \text{for all } t < 0.$$