随机过程

作者：Huang

时间：March 29, 2025

目录

第 1 章 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 $E_i \cap E_j = \emptyset$ when $i \neq j$,

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

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

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

1. If $E \subseteq F$, then $P(E) \leq P(F)$.
2. $P(E^c) = 1 - P(E)$ where E^c is the complement of E.
3. If the E_i are mutually exclusive, then
$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i).$$
4. In general,
$$P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i).$$

The inequatlity 4 is known as *Boole's inequatlity*

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

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

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

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

We may now state the following:

命题 1.1

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

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

♠

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

$$F_1 = E_1,$$
$$F_n = E_n \left(\bigcup_{i=1}^{n-1} E_i\right)^c = E_n E_{n-1}^c, \quad n > 1$$

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

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

Thus

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} E_i\right) &= P\left(\bigcup_{i=1}^{\infty} F_i\right) \\ &= \sum_{i=1}^{\infty} P(F_i) \quad \text{(by Axiom 3)} \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P(F_i) \\ &= \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} F_i\right) \\ &= \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} E_i\right) \\ &= \lim_{n\to\infty} P(E_n), \end{aligned}$$

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,

$$P\left(\bigcup_{i=1}^{\infty} E_i^c\right) = \lim_{n\to\infty} P(E_n^c)$$

But, as $\bigcup_{i=1}^{\infty} E_i^c = \left(\bigcap_{i=1}^{\infty} E_n\right)^c$, we see that

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

or, equivalently,

$$P\left(\bigcap_{i=1}^{\infty} E_n\right) = \lim_{n\to\infty} P(E_n),$$

which proves the result.

例题 1.1 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.

命题 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

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

♠

证明 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.

例题 1.2 Let $X_1, X_2, \ldots$ be such that

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

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.

命题 1.3 (Converse to the Borel-Cantelli Lemma)

If $E_1, E_2, \ldots$ are independent events such that

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

then

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

♠

证明

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

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

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

Now,

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

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

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

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

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

Hence the result follows.

例题 1.3 Let $X_1, X_2, \ldots$ be independent and such that

$$P\{X_n = 0\} = \frac{1}{n} = 1 - P\{X_n = 1\}, \quad n \geq 1$$

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 Varibles

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 Expercted Value

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

$$E[X] = \int_{-\infty}^{\infty} x \, dF(x) \tag{1.1}$$

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

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

$$E[h(X)] = \int_{-\infty}^{\infty} x \, dF_h(x), \tag{1.2}$$

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,

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) \, dF(x). \tag{1.3}$$

The variance of the random variable X is defined by

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

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

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

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

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]. \tag{1.4}$$

The corresponding property for variances is that

$$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i<j} \text{Cov}(X_i, X_j) \tag{1.5}$$

例题 1.4 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

$$E[X_i] = 1/n,$$
$$\text{Var}(X_i) = \frac{1}{n}\left(1 - \frac{1}{n}\right) = \frac{n-1}{n^2}$$

Also

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

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

$$E[X_i X_j] = P\{X_i = 1, X_j = 1\}$$
$$= P\{X_i = 1\}P\{X_j = 1 | X_i = 1\}$$
$$= \frac{1}{n} \cdot \frac{1}{n-1}.$$

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

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

Thus both the mean and variance of the number of matches are equal to 1.

例题 1.5 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 $A_i, 1 \le j \le n$, that occur. A useful identity can be obtained by noting that

$$(1-1)^N = \begin{cases} 1 & \text{if } N = 0 \\ 0 & \text{if } N > 0. \end{cases} \tag{1.6}$$

But by the binomial theorem,

$$(1-1)^n = \sum_{i=0}^{n} \binom{N}{i}(-1)^i \tag{1.7}$$

$$= \sum_{i=0}^{n} \binom{N}{i}(-1)^i \quad \text{since } \binom{m}{i} = 0 \text{ when } i > m.$$

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

$$I = \sum_{i=1}^{n} \binom{N}{i} (-1)^{i+1} \tag{1.8}$$

Taking expectations of both sides of (1.3.7) yields

$$E[I] = E[N] - E\left[\binom{N}{2}\right] + \cdots + (-1)^{n+1} E\left[\binom{N}{n}\right]. \tag{1.9}$$

However,

$$E[I] = P\{N > 0\}$$

$$= P\{\text{at least one of the } A_i \text{ occurs}\}$$

$$= P\left(\bigcup_{i=1}^{n} A_i\right)$$

and

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

$$E\left[\binom{N}{2}\right] = E[\text{number of pairs of the } A_i \text{ that occur}]$$

$$= E\left[\sum_{i<j} I_i I_j\right]$$

$$= \sum_{i<j} E[I_i I_j]$$

$$= \sum_{i<j} P(A_i A_j),$$

and, in general, by the same reasoning,

$$E\left[\binom{N}{i}\right] = E[\text{number of sets of size } i \text{ that occur}]$$

$$= E\left[\sum_{j_1<j_2<\cdots<j_i} I_{j_1} I_{j_2} \cdots I_{j_i}\right]$$

$$= \sum_{j_1<j_2<\cdots<j_i} P(A_{j_1} A_{j_2} \cdots A_{j_i}).$$

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

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

Other useful identities can also be derived by this approach. For instance, suppose we want a formula for the proba-

bility 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

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

Taking expectations of both sides of the above yields

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

$$P\{\text{exactly } r \text{ of the events } A_1, \ldots, A_n \text{ occur}\} \tag{1.10}$$

$$= \sum_{i=0}^{n-r}(-1)^i\binom{r+i}{r}\sum_{j_1<j_2<\cdots<j_{r+i}} P(A_{j_1}A_{j_2}\cdots A_{j_{r+i}}).$$

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

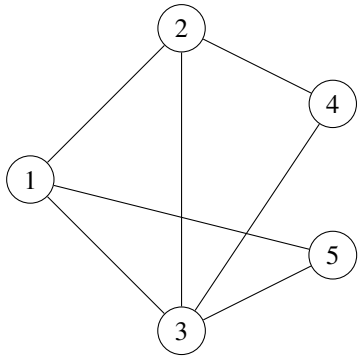
$$\begin{aligned} &P\{\text{exactly } r \text{ of the boxes are empty}\} \\ &= \sum_{i=0}^{n-r}(-1)^i\binom{r+i}{r}\binom{n}{r+i}\left(1-\frac{r+i}{n}\right)^m, \end{aligned}$$

where the above follows since $\sum_{j_1<j_2<\cdots<j_{r+i}}$ 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.

例题 1.6 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 $E = \{(1, 2), (1, 3), (1, 5), (2, 3), (2, 4), (3, 4), (3,$ 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 $A = \{1, 2, 4\}$).

解 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 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 Function,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,

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

Evaluating at $t = 0$ yields

$$\psi^{(n)}(0) = E[X^n], \quad n \geq 1$$

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.

例题 1.7 Let X and Y be independent normal random variables with respective means μ_1 and μ_2 and respective variances σ_1^2 and σ_2^2. The moment generating function of their sum is given by

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

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

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

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.

例题 1.8 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, $1 \le i \le m$,

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

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

Let us now consider

$$\psi(t_1, \ldots, t_m) = E\left[\exp\left\{\sum_{i=1}^{m} t_i X_i\right\}\right],$$

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

$$\text{Var}\left(\sum_{i=1}^{m} t_i X_i\right) = \text{Cov}\left(\sum_{i=1}^{m} t_i X_i, \sum_{j=1}^{m} t_j X_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} t_i t_j \text{Cov}(X_i, X_j).$$

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

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

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 $\text{Cov}(X_i, X_j)$, $i, j = 1, \ldots, m$.

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

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

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

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

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

$$E[X|Y = y] = \int x \, dF(x|y) = \sum_x x P\{X = x|Y = y\}$$

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

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

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

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

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

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

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

$$E[X] = E[E[X|Y]] = \int E[X|Y = y] dF_Y(y) \tag{1.11}$$

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

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

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.

例题 1.9 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 $Y = \sum_{i=1}^N X_i$ by first conditioning on N. Now

$$\begin{aligned}&E\left[\exp\left\{t\sum_{i=1}^N X_i\right\}\middle|N=n\right] \\ &= E\left[\exp\left\{t\sum_{i=1}^n X_i\right\}\middle|N=n\right] \\ &= E\left[\exp\left\{t\sum_{i=1}^n X_i\right\}\right] \quad \text{(by independence)} \\ &= (\psi_X(t))^n,\end{aligned}$$

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

$$E\left[\exp\left\{t\sum_{i=1}^N X_i\right\}\middle|N\right] = (\psi_X(t))^N$$

and so

$$\psi_Y(t) = E\left[\exp\left\{t\sum_{i=1}^N X_i\right\}\right] = E[(\psi_X(t))^N].$$

To compute the mean and variance of $Y = \sum_{i=1}^N X_i$, 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

$$E[Y^2] = E[N(N-1)E^2[X] + NE[X^2]]$$

$$= E[N]\text{Var}(X) + E^2[X]\text{Var}(N).$$

Hence,

$$\text{Var}(Y) = E[Y^2] - (E[Y])^2$$

$$= E[N]\text{Var}(X) + E^2[X]\text{Var}(N).$$

例题 1.10 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

$$E[e^{tX}] = \frac{1}{3}(E[e^{tX}|Y=1] + E[e^{tX}|Y=2] + E[e^{tX}|Y=3]). \tag{1.12}$$

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

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

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,

$$E[e^{tX}|Y=2] = E[e^{t(3+X')}] = e^{3t}E[e^{tX}].$$

Similarly,

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

Substitution back into (1.5.2) yields

$$E[e^{tX}] = \frac{1}{3}(e^{2t} + e^{3t}E[e^{tX}] + e^{5t}E[e^{tX}]) \tag{1.13}$$

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

$$E[X] = P(E)$$

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

Therefore, from Equation (1.5.1) we obtain that

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

1.5.1 Conditional Expectation 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 $X = (X_1, \ldots, X_n)$ 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 X a problem

of interest is to estimate the unseen value of θ. An estimator of θ can be any function $d(X)$ of the data, and in Bayesian statistics one often wants to choose $d(X)$ to minimize $E[(d(X) - \theta)^2|X]$, the conditional expected squared distance between the estimator and the parameter. Using the facts that

(i) conditional on X, $d(X)$ 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 $E[(d(X) - \theta)^2|X]$, called the Bayes estimator, is given by

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

An estimator $d(X)$ is said to be an unbiased estimator of θ if

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

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.

引理 1.1

For any random variable Y and random vector Z

$$E\left((Y - E[Y|Z])\, E[Y|Z]\right) = 0$$

♡

证明

$$E[Y\,E[Y|Z]] = E[E[Y\,E[Y|Z]|Z]]$$

$$= E[E[Y\,E[Y|Z]E[Y|Z]]$$

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

命题 1.4

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

♠

证明 Letting $Y = \theta$ and $Z = X$ in Lemma 1.5.1 yields that

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

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

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

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

$$E[(\theta - E[\theta|X])E[\theta|X]] + E[(E[\theta|X] - \theta)\theta] = 0 \tag{1.14}$$

or,

$$E[(\theta - E[\theta|X])E[\theta|X] + (E[\theta|X] - \theta)\theta] = 0 \tag{1.15}$$

or,

$$-E[(\theta - E[\theta|X])^2] = 0 \tag{1.16}$$

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

$$E[X] = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

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

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

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

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

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

例题 **1.11** 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 $\frac{1}{2}$.

例题 **1.12** 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

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

Hence, conditioning on R_i, yields

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

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 $\overline{F}(x) = P\{X > x\}$ Then $\overline{F}$ 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

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

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

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

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

$$P\{X \in (t, t+dt) \mid X > t\} = \frac{P\{X \in (t, t+dt), X > t\}}{P\{X > t\}} = \frac{P\{X \in (t, t+dt)\}}{P\{X > t\}} \approx \frac{f(t)\,dt}{\overline{F}(t)} = \lambda(t)\,dt.$$

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

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

Integration yields

$$\log \overline{F}(t) = -\int_0^t \lambda(\tau)\, d\tau + k$$

or

$$\overline{F}(t) = c \exp\left\{-\int_0^t \lambda(\tau)\, d\tau\right\}.$$

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

$$\overline{F}(t) = \exp\left\{-\int_0^t \lambda(\tau)\, d\tau\right\}.$$

引理 1.2

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

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

♡

证明 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

引理 1.3

If f is a convex function, then

$$E[f(X)] \geq f(E[X])$$

provided the expectations exist.

♡

证明 We will give a proof under the supposition that f has a Taylor series expansion. Expanding about $\mu = E[X]$ and using the Taylor series with a remainder formula yields

$$f(x) = f(\mu) + f'(\mu)(x-\mu) + f''(\xi)(x-\xi)^2/2 \geq f(\mu) + f'(\mu)(x-\mu)$$

since $f''(\xi) \geq 0$ by convexity Hence,

$$f(X) \geq f(\mu) + f'(\mu)(X-\mu)$$

Taking expectations gives that

$$E[f(X)] \geq f(\mu) + f'(\mu)E[X-\mu] = f(\mu).$$

1.7 Limit theorems

Some of the most important results in probability theory are in the form of limit theorems. The two most important are:

定理 1.1 (Strong Law of Large Numbers)

If $X_1, X_2, \ldots$ are independent and identically distributed with mean μ, then

$$P\left\{\lim_{n\to\infty} (X_1 + \cdots + X_n)/n = \mu\right\} = 1.$$

♡

定理 1.2 (Central Limit Theorem)

If $X_1, X_2, \ldots$ are independent and identically distributed with mean μ and variance σ^2, then

$$\lim_{n\to\infty} P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le a\right\} = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

♡

Thus if we let $S_n = \sum_{i=1}^{n} X_i$, where $X_1, X_2, \ldots$ are independent and identically distributed, then the Strong Law of Large Numbers states that, with probability 1, S_n/n will converge to $E[X_i]$; whereas the central limit theorem states that S_n will have an asymptotic normal distribution as $n \to \infty$.

1.8 Stochastic Process

A *stochastic process* $\mathbf{X} = \{X(t), t \in T\}$ is a collection of random variables. That is, for each t in the *index set* T, $X(t)$ is a random variable. We often interpret t as time and call $X(t)$ the state of the process at time t. If the index set T is a countable set, we call $\mathbf{X}$ a discrete-time stochastic process, and if T is a continuum, we call it a continuous-time process. Any realization of $\mathbf{X}$ is called a *sample path*. For instance, if events are occurring randomly in time and $X(t)$ represents

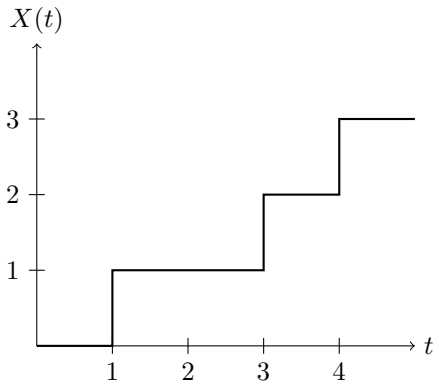


图 1.1: 1.9.1: A sample path of $X(t)$ = number of events in $[0, t]$.

the number of events that occur in $[0, t]$, then Figure 1.9.1 gives a sample path of $\mathbf{X}$ which corresponds to the initial event occurring at time 1, the next event at time 3 and the third at time 4, and no events anywhere else.

A continuous-time stochastic process $\{X(t), t \in T\}$ is said to have *independent increments* if for all $t_0 < t_1 < t_2 < \cdots < t_n$, the random variables

$$X(t_1) - X(t_0), X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1})$$

are independent. It is said to possess *stationary increments* if $X(t+s) - X(t)$ has the same distribution for all t. That is, it possesses independent increments if the changes in the process' value over nonoverlapping time intervals are independent; and it possesses stationary increments if the distribution of the change in value between any two points depends only on the distance between those points.

例题 1.13 Consider a particle that moves along a set of $m+1$ nodes, labelled $0, 1, \ldots, m$, that are arranged around a circle (see Figure 1.9.2) At each step the particle is equally likely to move one position in either the clockwise or counterclockwise direction. That is, if X_n is the position of the particle after its nth step then

$$P\{X_{n+1} = i + 1 \mid X_n = i\} = P\{X_{n+1} = i - 1 \mid X_n = i\} = \frac{1}{2}$$

where $i + 1 \equiv 0$ when $i = m$, and $i - 1 \equiv m$ when $i = 0$. Suppose now that the particle starts at 0 and continues to move around according to the above rules until all the nodes $1, 2, \ldots, m$ have been visited. What is the probability that node i, $i = 1, \ldots, m$, is the last one visited?

解 Surprisingly enough, the probability that node i is the last node visited can be determined without any computations.

To do so, consider the first time that the particle is at one of the two neighbors of node i, that is, the first time that the particle is at one of the nodes $i-1$ or $i+1$ (with $m+1 \equiv 0$). Suppose it is at node $i-1$ (the argument in the alternative situation is identical). Since neither node i nor $i+1$ has yet been visited it follows that i will be the last node visited if, and only if, $i+1$ is visited before i. This is so because in order to visit $i+1$ before i the particle will have to visit all the nodes on the counterclockwise path from $i-1$ to $i+1$ before it visits i. But the probability that a particle at node $i-1$ will visit $i+1$ before i is just the probability that a particle will progress $m-1$ steps in a specified direction before progressing one step in the other direction. That is, it is equal to the probability that a gambler who starts with 1 unit, and wins 1 when a fair coin turns up heads and loses 1 when it turns up tails, will have his fortune go up by $m-1$ before he goes broke. Hence, as the preceding implies that the probability that node i is the last node visited is the same for all i, and as these probabilities must sum to 1, we obtain

$$P\{i \text{ is the last node visited}\} = \frac{1}{m}, \quad i = 1, \ldots, m.$$

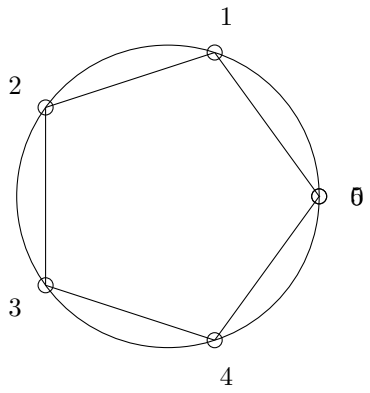


图 1.2: 1.9.2:Particle moving around a circle.

注 The argument used in the preceding example also shows that a gambler who is equally likely to either win or lose 1 on each gamble will be losing n before he is winning 1 with probability $1/(n+1)$, or equivalently

$$P\{\text{gambler is up 1 before being down } n\} = \frac{n}{n+1}.$$

Suppose now we want the probability that the gambler is up 2 before being down n. Upon conditioning upon whether he reaches up 1 before down n we obtain that

$$\begin{aligned} P\{\text{gambler is up 2 before being down } n\} &= P\{\text{up 2 before down } n \mid \text{up 1 before down } n\}\frac{n}{n+1} \\ &= P\{\text{up 1 before down } n+1\}\frac{n}{n+1} \\ &= \frac{n+1}{n+2}\frac{n}{n+1} = \frac{n}{n+2}. \end{aligned}$$

Repeating this argument yields that

$$P\{\text{gambler is up } k \text{ before being down } n\} = \frac{n}{n+k}.$$

例题 1.14 Suppose in Example 1.9(A) that the particle is not equally likely to move in either direction but rather moves at each step in the clockwise direction with probability p and in the counterclockwise direction with probability $q = 1-p$. If $0.5 < p < 1$ then we will show that the probability that state i is the last state visited is a strictly increasing function of i, $i = 1, \ldots, m$.

To determine the probability that state i is the last state visited, condition on whether $i-1$ or $i+1$ is visited first. Now, if $i-1$ is visited first then the probability that i will be the last state visited is the same as the probability that a gambler who wins each 1 unit bet with probability q will have her cumulative fortune increase by $m-1$ before it decreases by 1. Note that this probability does not depend on i, and let its value be P_1. Similarly, if $i+1$ is visited before $i-1$ then the

probability that i will be the last state visited is the same as the probability that a gambler who wins each 1 unit bet with probability p will have her cumulative fortune increase by $m-1$ before it decreases by 1. Call this probability P_2, and note that since $p>q$, $P_1<P_2$. Hence, we have

$$P\{i \text{ is last state}\} = P_1 P\{i-1 \text{ before } i+1\} + P_2(1 - P\{i-1 \text{ before } i+1\})$$

$$= (P_1 - P_2)P\{i-1 \text{ before } i+1\} + P_2$$

Now, since the event that $i-1$ is visited before $i+1$ implies the event that $i-2$ is visited before i, it follows that

$$P\{i-1 \text{ before } i+1\} < P\{i-2 \text{ before } i\},$$

and thus we can conclude that

$$P\{i-1 \text{ is last state}\} < P\{i \text{ is last state}\}$$

例题 1.15 A graph consisting of a central vertex, labeled 0, and rays emanating from that vertex is called a star graph (see Figure 1.9.3) Let r denote the number of rays of a star graph and let ray i consist of n_i vertices, for $i=1,\ldots,r$ Suppose that a particle moves along the vertices of the graph so that it is equally likely to move from whichever vertex it is presently at to any of the neighbors of that vertex, where two vertices are said to be neighbors if they are joined by an edge. Thus, for instance, when at vertex 0 the particle is equally likely to move to any of its r neighbors. The vertices at the far ends of the rays are called *leafs*. What is the probability that, starting at node 0, the first leaf visited is the one on ray i, $i=1,\ldots,r$?

解 Let L denote the first leaf visited. Conditioning on R, the first ray visited, yields

$$P\{L=i\} = \sum_{j=1}^{r} \frac{1}{r} P\{L=i \mid \text{first ray visited is } j\} \tag{1.17}$$

Now, if j is the first ray visited (that is, the first move of the particle is from vertex 0 to its neighboring vertex on ray j) then it follows, from the remark following Example 1.9(A), that with probability $1/n_j$ the particle will visit the leaf at the end of ray j before returning to 0 (for this is the complement of the event that the gambler will be up 1 before being down $n-1$) Also, if it does return to 0 before reaching the end of ray j, then the problem in essence begins anew Hence, we obtain upon conditioning whether the particle reaches the end of ray j before returning to 0 that

$$P\{L=i \mid \text{first ray visited is } i\} = \frac{1}{n_i} + (1 - \frac{1}{n_i})P\{L=i\}$$

$$P\{L=i \mid \text{firstrayvisitedis}j\} = (1 - \frac{1}{n_j})P\{L=i\}, \quad \text{for}j \neq i$$

Substituting the preceding into Equation (191) yields that

$$rP\{L=i\} = \frac{1}{n_i} + \left(r - \sum_j \frac{1}{n_j}\right) P\{L=i\}$$

or

$$P\{L=i\} = \frac{\frac{1}{n_i}}{\sum_j \frac{1}{n_j}}, \quad i=1,\ldots,r.$$

第 2 章 The Poisson Process

2.1 The Poisson Process

A stochastic process $\{N(t),\, t \geq 0\}$ is said to be a **counting process** if $N(t)$ represents the total number of 'events' that have occurred up to time t. Hence, a counting process $N(t)$ must satisfy:

(i) $N(t) \geq 0$.

(ii) $N(t)$ is integer valued.

(iii) If $s < t$, then $N(s) \leq N(t)$.

(iv) For $s < t$, $N(t) - N(s)$ equals the number of events that have occurred in the interval $(s, t]$.

A counting process is said to possess **independent increments** if the numbers of events that occur in disjoint time intervals are independent. For example, this means that the number of events that have occurred by time t (that is, $N(t)$) must be independent of the number of events occurring between times t and $t + s$ (that is, $N(t + s) - N(t)$).

A counting process is said to possess **stationary increments** if the distribution of the number of events that occur in any interval of time depends only on the length of the time interval. In other words, the process has stationary increments if the number of events in the interval $(t_1 + s, t_2 + s]$ (that is, $N(t_2 + s) - N(t_1 + s)$) has the same distribution as the number of events in the interval $(t_1, t_2]$ (that is, $N(t_2) - N(t_1)$) for all $t_1 \leq t_2$, and $s > 0$.

One of the most important types of counting processes is the **Poisson process**, which is defined as follows.

定义 2.1

The counting process $\{N(t), t \geq 0\}$ is said to be a *Poisson process* having rate λ, $\lambda > 0$, if

(i) $N(0) = 0$

(ii) The process has independent increments

(iii) The number of events in any interval of length t is Poisson distributed with mean λt. That is, for all $s, t \geq 0$,

$$P\{N(t + s) - N(s) = n\} = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \quad n = 0, 1, \tag{2.1}$$

♣

Note that it follows from condition (iii) that a Poisson process has stationary increments and also that

$$E[N(t)] = \lambda t,$$

which explains why λ is called the rate of the process.

In order to determine if an arbitrary counting process is actually a Poisson process, we must show that conditions (i), (ii), and (iii) are satisfied. Condition (i), which simply states that the counting of events begins at time $t = 0$, and condition (ii) can usually be directly verified from our knowledge of the process. However, it is not at all clear how we would determine that condition (iii) is satisfied, and for this reason an equivalent definition of a Poisson process would be useful.

As a prelude to giving a second definition of a Poisson process, we shall define the concept of a function f being $o(h)$.

定义 2.2

The function f is said to be $o(h)$ if

$$\lim_{h \to 0} \frac{f(h)}{h} = 0.$$

♣

We are now in a position to give an alternative definition of a Poisson process.

定义 2.3

The counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate λ, $\lambda > 0$, if

(i) $N(0) = 0$

(ii) The process has stationary and independent increments.

(iii) $P\{N(h) = 1\} = \lambda h + o(h)$

(iv) $P\{N(h) \geq 2\} = o(h)$

♣

定理 2.1

Two definitions are equivalent

♡

证明 We first show that Definition 2.12 implies Definition 2.11 To this let

$$P_n(t) = P\{N(t) = n\}$$

We derive a differential equation for $P_n(t)$ in the following manner

$$P_0(t+h) = P\{N(t+h) = 0\}$$

$$= P\{N(t) = 0, N(t+h) - N(t) = 0\}$$

$$= P\{N(t) = 0\}P\{N(t+h) - N(t) = 0\}$$

$$= P_0(0)[1 - \lambda h + o(h)],$$

where the final two equations follow from Assumption (ii) and the fact that (iii) and (iv) imply that $P\{N(h) = 0\} = 1 - \lambda h + o(h)$. Hence,

$$\frac{P_n(t+h) - P_n(t)}{h} = -\lambda P_n(t) + \frac{o(h)}{h}.$$

Letting $h \to 0$ yields

$$P_n'(t) = -\lambda P_n(t)$$

or

$$\frac{P_n'(t)}{P_n(t)} = -\lambda,$$

which implies, by integration,

$$\log P_n(t) = -\lambda t + c$$

or

$$P_n(t) = Ke^{-\lambda t}.$$

Since $P_n(0) = P\{N(0) = 0\} = 1$, we arrive at

$$P_n(t) = e^{-\lambda t}. \tag{2.2}$$

Similarly, for $n \geq 1$,

$$P_n(t+h) = P\{N(t+h) = n\}$$

$$= P\{N(t) = n, N(t+h) - N(t) = 0\}$$

$$+P\{N(t) = n-1, N(t+h) - N(t) = 1\}$$

$$+P\{N(t+h) = n, N(t+h) - N(t) \geq 2\}$$

However, by (iv), the last term in the above is $o(h)$, hence, by using (ii), we obtain

$$P_n(t+h) = P_n(t)P_0(h) + P_{n-1}(t)P_1(h) + o(h)$$

$$= (1 - \lambda h)P_n(t) + \lambda h P_{n-1}(t) + o(h)$$

Thus,

$$\frac{P_n(t+h) - P_n(t)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(h)}{h}$$

Letting $h \to 0$,

$$P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t),$$

or, equivalently,

$$e^{\lambda t}[P_n'(t) + \lambda P_n(t)] = \lambda e^{\lambda t} P_{n-1}(t)$$

Hence,

$$\frac{d}{dt}(e^{\lambda t} P_n(t)) = \lambda e^{\lambda t} P_{n-1}(t) \tag{2.3}$$

Now by (2.11) we have when $n = 1$

$$\frac{d}{dt}(e^{\lambda t} P_1(t)) = \lambda$$

or

$$P_1(t) = (\lambda t + c)e^{-\lambda t},$$

which, since $P_1(0) = 0$, yields

$$P_1(t) = \lambda t e^{-\lambda t}$$

To show that $P_n(t) = e^{-\lambda t}(\lambda t)^n / n!$, we use mathematical induction and hence first assume it for $n - 1$ Then by (2.12),

$$\frac{d}{dt}(e^{\lambda t} P_n(t)) = \frac{\lambda(\lambda t)^{n-1}}{(n-1)!}$$

implying that

$$e^{\lambda t} P_n(t) = \frac{(\lambda t)^n}{n!} + c,$$

or, since $P_n(0) = P\{N(0) = n\} = 0$,

$$P_n(t) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

Thus Definition 2.12 implies Definition 2.11 We will leave it for the reader to prove the reverse

注

The result that $N(t)$ has a Poisson distribution is a consequence of the Poisson approximation to the binomial distribution. To see this subdivide the interval $[0, t]$ into k equal parts where k is very large (Figure 2.11). First we note that the probability of having 2 or more events in any subinterval goes to 0 as $k \to \infty$. This follows from

$$P\{2 \text{ or more events in any subinterval}\}$$
$$\leq \sum_{i=1}^{k} P\{2 \text{ or more events in the } i\text{th subinterval}\}$$
$$= ko\left(\frac{t}{k}\right)$$
$$= t\frac{o(t/k)}{t/k}$$
$$\to 0 \quad \text{as } k \to \infty.$$

Hence, $N(t)$ will (with a probability going to 1) just equal the number of subintervals in which an event occurs. However, by stationary and independent increments this number will have a binomial distribution with parameters k and $p = \lambda t/k + o(t/k)$. Hence by the Poisson approximation to the binomial we see by letting k approach ∞ that $N(t)$ will

have a Poisson distribution with mean equal to

$$\lim_{k\to\infty} k\left[\lambda\frac{t}{k}+o\left(\frac{t}{k}\right)\right]=\lambda t+\lim_{k\to\infty}\left[t\frac{o(t/k)}{t/k}\right]=\lambda t.$$

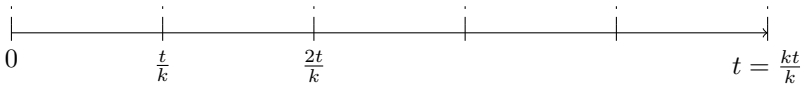


Figure 2.1.1

图 2.1: 将区间 $[0,t]$ 划分为 k 个相等的子区间