Elements of Probability

2.1 Sample Space and Events

Consider an experiment whose outcome is not known in advance. Let S, called the sample space of the experiment, denote the set of all possible outcomes. For example, if the experiment consists of the running of a race among the seven horses numbered 1 through 7, then

$$S = \{\text{all orderings of } (1, 2, 3, 4, 5, 6, 7)\}$$

The outcome (3, 4, 1, 7, 6, 5, 2) means, for example, that the number 3 horse came in first, the number 4 horse came in second, and so on.

Any subset A of the sample space is known as an event. That is, an event is a set consisting of possible outcomes of the experiment. If the outcome of the experiment is contained in A, we say that A has occurred. For example, in the above, if

$$A = \{\text{all outcomes in } S \text{ starting with } 5\}$$

then A is the event that the number 5 horse comes in first.

For any two events A and B we define the new event $A \cup B$, called the union of A and B, to consist of all outcomes that are either in A or B or in both A and B. Similarly, we define the event AB, called the intersection of A and B, to consist of all outcomes that are in both A and B. That is, the event $A \cup B$ occurs if either A or B occurs, whereas the event AB occurs if both A and B occur. We can also define unions and intersections of more than two events. In particular, the union of the events $A_1, \ldots, A_n$—designated by $\cup_{i=1}^{n} A_i$—is defined to consist of all outcomes that are in any of the A_i. Similarly, the intersection of the events $A_1, \ldots, A_n$—designated by $A_1 A_2 \cdots A_n$—is defined to consist of all outcomes that are in all of the A_i.

Simulation. DOI: http://dx.doi.org/10.1016/B978-0-12-415825-2.00002-4

For any event A we define the event A^c, referred to as the complement of A, to consist of all outcomes in the sample space S that are not in A. That is, A^c occurs if and only if A does not. Since the outcome of the experiment must lie in the sample space S, it follows that S^c does not contain any outcomes and thus cannot occur. We call S^c the null set and designate it by $\emptyset$. If $AB = \emptyset$ so that A and B cannot both occur (since there are no outcomes that are in both A and B), we say that A and B are mutually exclusive.

2.2 Axioms of Probability

Suppose that for each event A of an experiment having sample space S there is a number, denoted by $P(A)$ and called the probability of the event A, which is in accord with the following three axioms:

Axiom 1 $0 \leqslant P(A) \leqslant 1$

Axiom 2 $P(S) = 1$

Axiom 3 *For any sequence of mutually exclusive events* $A_1, A_2, \ldots$

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i), \quad n = 1, 2, \ldots, \infty.$$

Thus, Axiom 1 states that the probability that the outcome of the experiment lies within A is some number between 0 and 1; Axiom 2 states that with probability 1 this outcome is a member of the sample space; and Axiom 3 states that for any set of mutually exclusive events, the probability that at least one of these events occurs is equal to the sum of their respective probabilities.

These three axioms can be used to prove a variety of results about probabilities. For instance, since A and A^c are always mutually exclusive, and since $A \cup A^c = S$, we have from Axioms 2 and 3 that

$$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$$

or equivalently

$$P(A^c) = 1 - P(A)$$

In words, the probability that an event does not occur is 1 minus the probability that it does.

2.3 Conditional Probability and Independence

Consider an experiment that consists of flipping a coin twice, noting each time whether the result was heads or tails. The sample space of this experiment can be taken to be the following set of four outcomes:

$$S = \{(\text{H, H}), (\text{H, T}), (\text{T, H}), (\text{T, T})\}$$

where (H, T) means, for example, that the first flip lands heads and the second tails. Suppose now that each of the four possible outcomes is equally likely to occur and thus has probability $\frac{1}{4}$. Suppose further that we observe that the first flip lands on heads. Then, given this information, what is the probability that both flips land on heads? To calculate this probability we reason as follows: Given that the initial flip lands heads, there can be at most two possible outcomes of our experiment, namely, (H, H) or (H, T). In addition, as each of these outcomes originally had the same probability of occurring, they should still have equal probabilities. That is, given that the first flip lands heads, the (conditional) probability of each of the outcomes (H, H) and (H, T) is $\frac{1}{2}$, whereas the (conditional) probability of the other two outcomes is 0. Hence the desired probability is $\frac{1}{2}$.

If we let A and B denote, respectively, the event that both flips land on heads and the event that the first flip lands on heads, then the probability obtained above is called the conditional probability of A given that B has occurred and is denoted by

$$P(A|B)$$

A general formula for $P(A|B)$ that is valid for all experiments and events A and B can be obtained in the same manner as given previously. Namely, if the event B occurs, then in order for A to occur it is necessary that the actual occurrence be a point in both A and B; that is, it must be in AB. Now since we know that B has occurred, it follows that B becomes our new sample space and hence the probability that the event AB occurs will equal the probability of AB relative to the probability of B. That is,

$$P(A|B) = \frac{P(AB)}{P(B)}.$$

The determination of the probability that some event A occurs is often simplified by considering a second event B and then determining both the conditional probability of A given that B occurs and the conditional probability of A given that B does not occur. To do this, note first that

$$A = AB \cup AB^c.$$

Because AB and AB^c are mutually exclusive, the preceding yields

$$\begin{aligned} P(A) &= P(AB) + P(AB^c) \\ &= P(A|B)P(B) + P(A|B^c)P(B^c) \end{aligned}$$

When we utilize the preceding formula, we say that we are computing $P(A)$ by *conditioning on whether or not B occurs*.

Example 2a An insurance company classifies its policy holders as being either accident prone or not. Their data indicate that an accident prone person will file a claim within a one-year period with probability .25, with this probability falling to .10 for a non accident prone person. If a new policy holder is accident prone with probability .4, what is the probability he or she will file a claim within a year?

Solution Let C be the event that a claim will be filed, and let B be the event that the policy holder is accident prone. Then

$$P(C) = P(C|B)P(B) + P(C|B^c)P(B^c) = (.25)(.4) + (.10)(.6) = .16 \qquad \square$$

Suppose that exactly one of the events $B_i, i = 1, \ldots, n$ must occur. That is, suppose that $B_1, B_2, \ldots, B_n$ are mutually exclusive events whose union is the sample space S. Then we can also compute the probability of an event A by conditioning on which of the B_i occur. The formula for this is obtained by using that

$$A = AS = A(\cup_{i=1}^{n} B_i) = \cup_{i=1}^{n} AB_i$$

which implies that

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(AB_i) \\ &= \sum_{i=1}^{n} P(A|B_i)P(B_i) \end{aligned}$$

Example 2b Suppose there are k types of coupons, and that each new one collected is, independent of previous ones, a type j coupon with probability p_j, $\sum_{j=1}^{k} p_j = 1$. Find the probability that the n^{th} coupon collected is a different type than any of the preceding $n - 1$.

Solution Let N be the event that coupon n is a new type. To compute $P(N)$, condition on which type of coupon it is. That is, with T_j being the event that coupon n is a type j coupon, we have

$$\begin{aligned} P(N) &= \sum_{j=1}^{k} P(N|T_j)P(T_j) \\ &= \sum_{j=1}^{k} (1 - p_j)^{n-1} p_j \end{aligned}$$

where $P(N|T_j)$ was computed by noting that the conditional probability that coupon n is a new type given that it is a type j coupon is equal to the conditional probability that each of the first $n-1$ coupons is not a type j coupon, which by independence is equal to $(1-p_j)^{n-1}$. □

As indicated by the coin flip example, $P(A|B)$, the conditional probability of A, given that B occurred, is not generally equal to $P(A)$, the unconditional probability of A. In other words, knowing that B has occurred generally changes the probability that A occurs (what if they were mutually exclusive?). In the special case where $P(A|B)$ is equal to $P(A)$, we say that A and B are independent. Since $P(A|B) = P(AB)/P(B)$, we see that A is independent of B if

$$P(AB) = P(A)P(B)$$

Since this relation is symmetric in A and B, it follows that whenever A is independent of B, B is independent of A.

2.4 Random Variables

When an experiment is performed we are sometimes primarily concerned about the value of some numerical quantity determined by the result. These quantities of interest that are determined by the results of the experiment are known as random variables.

The cumulative distribution function, or more simply the distribution function, F of the random variable X is defined for any real number x by

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

A random variable that can take either a finite or at most a countable number of possible values is said to be discrete. For a discrete random variable X we define its probability mass function $p(x)$ by

$$p(x) = P\{X = x\}$$

If X is a discrete random variable that takes on one of the possible values $x_1, x_2, \ldots$, then, since X must take on one of these values, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1.$$

Example 2a Suppose that X takes on one of the values 1, 2, or 3. If

$$p(1) = \frac{1}{4}, \quad p(2) = \frac{1}{3}$$

then, since $p(1) + p(2) + p(3) = 1$, it follows that $p(3) = \frac{5}{12}$. □

Whereas a discrete random variable assumes at most a countable set of possible values, we often have to consider random variables whose set of possible values is an interval. We say that the random variable X is a continuous random variable if there is a nonnegative function $f(x)$ defined for all real numbers x and having the property that for any set C of real numbers

$$P\{X \in C\} = \int_C f(x)dx \tag{2.1}$$

The function f is called the probability density function of the random variable X.

The relationship between the cumulative distribution $F(\cdot)$ and the probability density $f(\cdot)$ is expressed by

$$F(a) = P\{X \in (-\infty, a)\} = \int_{-\infty}^{a} f(x)dx.$$

Differentiating both sides yields

$$\frac{d}{da}F(a) = f(a).$$

That is, the density is the derivative of the cumulative distribution function. A somewhat more intuitive interpretation of the density function may be obtained from Eqution (2.1) as follows:

$$P\left\{a - \frac{\epsilon}{2} \leqslant X \leqslant a + \frac{\epsilon}{2}\right\} = \int_{a-\epsilon/2}^{a+\epsilon/2} f(x)dx \approx \epsilon f(a)$$

when ϵ is small. In other words, the probability that X will be contained in an interval of length ϵ around the point a is approximately $\epsilon f(a)$. From this, we see that $f(a)$ is a measure of how likely it is that the random variable will be near a.

In many experiments we are interested not only in probability distribution functions of individual random variables, but also in the relationships between two or more of them. In order to specify the relationship between two random variables, we define the joint cumulative probability distribution function of X and Y by

$$F(x, y) = P\{X \leqslant x, \quad Y \leqslant y\}$$

Thus, $F(x, y)$ specifies the probability that X is less than or equal to x and simultaneously Y is less than or equal to y.

If X and Y are both discrete random variables, then we define the joint probability mass function of X and Y by

$$p(x, y) = P\{X = x, \quad Y = y\}$$

Similarly, we say that X and Y are jointly continuous, with joint probability density function $f(x, y)$, if for any sets of real numbers C and D

$$P\{X \in C, Y \in D\} = \iint_{\substack{x \in C \\ y \in D}} f(x, y)dx\, dy$$

The random variables X and Y are said to be independent if for any two sets of real numbers C and D

$$P\{X \in C, Y \in D\} = P\{X \in C\}P\{Y \in D\}.$$

That is, X and Y are independent if for all sets C and D the events $A = \{X \in C\}$ and $B = \{Y \in D\}$ are independent. Loosely speaking, X and Y are independent if knowing the value of one of them does not affect the probability distribution of the other. Random variables that are not independent are said to be dependent.

Using the axioms of probability, we can show that the discrete random variables X and Y will be independent if and only if , for all x, y,

$$P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\}$$

Similarly, if X and Y are jointly continuous with density function $f(x, y)$, then they will be independent if and only if, for all x, y,

$$f(x, y) = f_X(x) f_Y(y)$$

where $f_X(x)$ and $f_Y(y)$ are the density functions of X and Y, respectively.

2.5 Expectation

One of the most useful concepts in probability is that of the expectation of a random variable. If X is a discrete random variable that takes on one of the possible values $x_1, x_2, \ldots$, then the *expectation* or *expected value* of X, also called the mean of X and denoted by $E[X]$, is defined by

$$E[X] = \sum_i x_i P\{X = x_i\} \tag{2.2}$$

In words, the expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes it. For example, if the probability mass function of X is given by

$$p(0) = \frac{1}{2} = p(1)$$

then

$$E[X] = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = \frac{1}{2}$$

is just the ordinary average of the two possible values 0 and 1 that X can assume. On the other hand, if

$$p(0) = \frac{1}{3}, \qquad p(1) = \frac{2}{3}$$

then

$$E[X] = 0\left(\frac{1}{3}\right) + 1\left(\frac{2}{3}\right) = \frac{2}{3}$$

is a weighted average of the two possible values 0 and 1 where the value 1 is given twice as much weight as the value 0 since $p(1) = 2p(0)$.

Example 2b If I is an indicator random variable for the event A, that is, if

$$I = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur} \end{cases}$$

then

$$E[I] = 1P(A) + 0P(A^c) = P(A)$$

Hence, the expectation of the indicator random variable for the event A is just the probability that A occurs. □

If X is a continuous random variable having probability density function f, then, analogous to Equation (2.2), we define the expected value of X by

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

Example 2c If the probability density function of X is given by

$$f(x) = \begin{cases} 3x^2 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

then

$$E[X] = \int_0^1 3x^3 dx = \frac{3}{4}.$$ □

Suppose now that we wanted to determine the expected value not of the random variable X but of the random variable $g(X)$, where g is some given function. Since $g(X)$ takes on the value $g(x)$ when X takes on the value x, it seems intuitive that $E[g(X)]$ should be a weighted average of the possible values $g(x)$ with, for a given x, the weight given to $g(x)$ being equal to the probability (or probability density in the continuous case) that X will equal x. Indeed, the preceding can be shown to be true and we thus have the following result.

Proposition *If X is a discrete random variable having probability mass function $p(x)$, then*

$$E[g(X)] = \sum_x g(x)p(x)$$

whereas if X is continuous with probability density function $f(x)$, then

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

A consequence of the above proposition is the following.

Corollary *If a and b are constants*, then

$$E[aX + b] = aE[X] + b$$

Proof In the discrete case

$$\begin{aligned} E[aX + b] &= \sum_x (ax + b)p(x) \\ &= a\sum_x xp(x) + b\sum_x p(x) \\ &= aE[X] + b \end{aligned}$$

Since the proof in the continuous case is similar, the result is established. □

It can be shown that expectation is a linear operation in the sense that for any two random variables X_1 and X_2

$$E[X_1 + X_2] = E[X_1] + E[X_2]$$

which easily generalizes to give

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

2.6 Variance

Whereas $E[X]$, the expected value of the random variable X, is a weighted average of the possible values of X, it yields no information about the variation of these values. One way of measuring this variation is to consider the average value of the square of the difference between X and $E[X]$. We are thus led to the following definition.

Definition *If X is a random variable with mean μ, then the variance of X, denoted by $\mathrm{Var}(X)$, is defined by*

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

An alternative formula for $\mathrm{Var}(X)$ is derived as follows:

$$\begin{aligned}\mathrm{Var}(X) &= E\left[(X-\mu)^2\right]\\ &= E\left[X^2 - 2\mu X + \mu^2\right]\\ &= E\left[X^2\right] - E[2\mu X] + E\left[\mu^2\right]\\ &= E\left[X^2\right] - 2\mu E[X] + \mu^2\\ &= E\left[X^2\right] - \mu^2\end{aligned}$$

That is,

$$\mathrm{Var}(X) = E\left[X^2\right] - (E[X])^2$$

A useful identity, whose proof is left as an exercise, is that for any constants a and b

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$$

Whereas the expected value of a sum of random variables is equal to the sum of the expectations, the corresponding result is not, in general, true for variances. It is, however, true in the important special case where the random variables are independent. Before proving this let us define the concept of the covariance between two random variables.

Definition *The covariance of two random variables X and Y, denoted $\mathrm{Cov}(X, Y)$, is defined by*

$$\mathrm{Cov}(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right]$$

where $\mu_x = E[X]$ and $\mu_y = E[Y]$.

A useful expression for $\text{Cov}(X, Y)$ is obtained by expanding the right side of the above equation and then making use of the linearity of expectation. This yields

$$\begin{aligned}\text{Cov}(X, Y) &= E\left[XY - \mu_x Y - X\mu_y + \mu_x\mu_y\right] \\ &= E[XY] - \mu_x E[Y] - E[X]\mu_y + \mu_x\mu_y \\ &= E[XY] - E[X]E[Y] \end{aligned} \tag{2.3}$$

We now derive an expression for $\text{Var}(X+Y)$ in terms of their individual variances and the covariance between them. Since

$$E[X+Y] = E[X] + E[Y] = \mu_x + \mu_y$$

we see that

$$\begin{aligned}\text{Var}(X+Y) &= E\left[(X+Y-\mu_x-\mu_y)^2\right] \\ &= E\left[(X-\mu_x)^2 + (Y-\mu_y)^2 + 2(X-\mu_x)(Y-\mu_y)\right] \\ &= E\left[(X-\mu_x)^2\right] + E\left[(Y-\mu_y)^2\right] + 2E\left[(X-\mu_x)(Y-\mu_y)\right] \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) \end{aligned} \tag{2.4}$$

We end this section by showing that the variance of the sum of independent random variables is equal to the sum of their variances.

Proposition *If X and Y are independent random variables then*

$$\text{Cov}(X, Y) = 0$$

and so, from Equation (2.4),

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

Proof From Equation (2.3) it follows that we need to show that $E[XY] = E[X]E[Y]$. Now in the discrete case,

$$\begin{aligned}E[XY] &= \sum_j \sum_i x_i y_j P\{X = x_i, Y = y_j\} \\ &= \sum_j \sum_i x_i y_j P\{X = x_i\} P\{Y = y_j\} \quad \text{by independence} \\ &= \sum_j y_j P\{Y = y_j\} \sum_i x_i P\{X = x_i\} \\ &= E[Y]E[X] \end{aligned}$$

Since a similar argument holds in the continuous case, the result is proved. □

The *correlation* between two random variables X and Y, denoted as $\text{Corr}(X, Y)$, is defined by

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

2.7 Chebyshev's Inequality and the Laws of Large Numbers

We start with a result known as Markov's inequality.

Proposition Markov's Inequality *If X takes on only nonnegative values, then for any value $a > 0$*

$$P\{X \geqslant a\} \leqslant \frac{E[X]}{a}$$

Proof Define the random variable Y by

$$Y = \begin{cases} a, & \text{if } X \geqslant a \\ 0, & \text{if } X < a \end{cases}$$

Because $X \geqslant 0$, it easily follows that

$$X \geqslant Y$$

Taking expectations of the preceding inequality yields

$$E[X] \geqslant E[Y] = aP\{X \geqslant a\}$$

and the result is proved. □

As a corollary we have Chebyshev's inequality, which states that the probability that a random variable differs from its mean by more than k of its standard deviations is bounded by $1/k^2$, where the standard deviation of a random variable is defined to be the square root of its variance.

Corollary Chebyshev's Inequality *If X is a random variable having mean μ and variance σ^2, then for any value $k > 0$,*

$$P\{|X - \mu| \geqslant k\sigma\} \leqslant \frac{1}{k^2}$$

Proof Since $(X - \mu)^2/\sigma^2$ is a nonnegative random variable whose mean is

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

we obtain from Markov's inequality that

$$P\left\{\frac{(X-\mu)^2}{\sigma^2} \geqslant k^2\right\} \leqslant \frac{1}{k^2}$$

The result now follows since the inequality $(X - \mu)^2/\sigma^2 \geqslant k^2$ is equivalent to the inequality $|X - \mu| \geqslant k\sigma$. □

We now use Chebyshev's inequality to prove the weak law of large numbers, which states that the probability that the average of the first n terms of a sequence of independent and identically distributed random variables differs from its mean by more than ϵ goes to 0 as n goes to infinity.

Theorem The Weak Law of Large Numbers *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables having mean μ. Then, for any $\epsilon > 0$,*

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} \to 0 \quad \text{as } n \to \infty$$

Proof We give a proof under the additional assumption that the random variables X_i have a finite variance σ^2. Now

$$E\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n}(E[X_1] + \cdots + E[X_n]) = \mu$$

and

$$\text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}[\text{Var}(X_1) + \cdots + \text{Var}(X_n)] = \frac{\sigma^2}{n}$$

where the above equation makes use of the fact that the variance of the sum of independent random variables is equal to the sum of their variances. Hence, from Chebyshev's inequality, it follows that for any positive k

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geqslant \frac{k\sigma}{\sqrt{n}}\right\} \leqslant \frac{1}{k^2}$$

Hence, for any $\epsilon > 0$, by letting k be such that $k\sigma/\sqrt{n} = \epsilon$, that is, by letting $k^2 = n\epsilon^2/\sigma^2$, we see that

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geqslant \epsilon\right\} \leqslant \frac{\sigma^2}{n\epsilon^2}$$

which establishes the result. □

A generalization of the weak law is the strong law of large numbers, which states that, with probability 1,

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu$$

That is, with certainty, the long-run average of a sequence of independent and identically distributed random variables will converge to its mean.

2.8 Some Discrete Random Variables

There are certain types of random variables that frequently appear in applications. In this section we survey some of the discrete ones.

Binomial Random Variables

Suppose that n independent trials, each of which results in a "success" with probability p, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to be a binomial random variable with parameters (n, p). Its probability mass function is given by

$$P_i \equiv P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n \tag{2.5}$$

where

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}$$

is the binomial coefficient, equal to the number of different subsets of i elements that can be chosen from a set of n elements.

The validity of Equation (2.5) can be seen by first noting that the probability of any particular sequence of outcomes that results in i successes and $n-i$ failures is, by the assumed independence of trials, $p^i(1-p)^{n-i}$. Equation (2.5) then follows since there are $\binom{n}{i}$ different sequences of the n outcomes that result in i successes and $n-i$ failures—which can be seen by noting that there are $\binom{n}{i}$ different choices of the i trials that result in successes.

A binomial $(1, p)$ random variable is called a Bernoulli random variable. Since a binomial (n, p) random variable X represents the number of successes in n independent trials, each of which results in a success with probability p, we can represent it as follows:

$$X = \sum_{i=1}^{n} X_i \tag{2.6}$$

where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a success} \\ 0 & \text{otherwise} \end{cases}$$

Now

$$\begin{aligned} E[X_i] &= P\{X_i = 1\} = p \\ \text{Var}(X_i) &= E\left[X_i^2\right] - E([X_i])^2 \\ &= p - p^2 = p(1-p) \end{aligned}$$

where the above equation uses the fact that $X_i^2 = X_i$ (since $0^2 = 0$ and $1^2 = 1$). Hence the representation (2.6) yields that, for a binomial (n, p) random variable X,

$$\begin{aligned} E[X] &= \sum_{i=1}^{n} E[X_i] = np \\ \mathrm{Var}(X) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \quad \text{since the } X_i \text{ are independent} \\ &= np(1-p) \end{aligned}$$

The following recursive formula expressing p_{i+1} in terms of p_i is useful when computing the binomial probabilities:

$$\begin{aligned} p_{i+1} &= \frac{n!}{(n-i-1)!(i+1)!} p^{i+1}(1-p)^{n-i-1} \\ &= \frac{n!(n-i)}{(n-i)!i!(i+1)} p^i (1-p)^{n-i} \frac{p}{1-p} \\ &= \frac{n-i}{i+1} \frac{p}{1-p} p_i \end{aligned}$$

Poisson Random Variables

A random variable X that takes on one of the values $0, 1, 2, \ldots$ is said to be a Poisson random variable with parameter λ, $\lambda > 0$, if its probability mass function is given by

$$p_i = P\{X = i\} = e^{-\lambda} \frac{\lambda^i}{i!}, \quad i = 0, 1, \ldots$$

The symbol e, defined by $e = \lim_{n \to \infty} (1 + 1/n)^n$, is a famous constant in mathematics that is roughly equal to 2.7183.

Poisson random variables have a wide range of applications. One reason for this is that such random variables may be used to approximate the distribution of the number of successes in a large number of trials (which are either independent or at most "weakly dependent") when each trial has a small probability of being a success. To see why this is so, suppose that X is a binomial random variable with parameters (n, p)—and so represents the number of successes in n independent trials when each trial is a success with probability p—and let $\lambda = np$. Then

$$\begin{aligned} P\{X = i\} &= \frac{n!}{(n-i)!i!} p^i (1-p)^{n-i} \\ &= \frac{n!}{(n-i)!i!} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i} \end{aligned}$$

Now for n large and p small,

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \qquad \frac{n(n-1)\cdots(n-i+1)}{n^i} \approx 1, \qquad \left(1-\frac{\lambda}{n}\right)^i \approx 1$$

Hence, for n large and p small,

$$P\{X=i\} \approx e^{-\lambda}\frac{\lambda^i}{i!}$$

Since the mean and variance of a binomial random variable Y are given by

$$E[Y] = np, \qquad \text{Var}(Y) = np(1-p) \approx np \quad \text{for small } p$$

it is intuitive, given the relationship between binomial and Poisson random variables, that for a Poisson random variable, X, having parameter λ,

$$E[X] = \text{Var}(X) = \lambda$$

An analytic proof of the above is left as an exercise.

To compute the Poisson probabilities we make use of the following recursive formula:

$$\frac{p_{i+1}}{p_i} = \frac{\frac{e^{-\lambda}\lambda^{i+1}}{(i+1)!}}{\frac{e^{-\lambda}\lambda^i}{i!}} = \frac{\lambda}{i+1}$$

or, equivalently,

$$p_{i+1} = \frac{\lambda}{i+1}p_i, \quad i \geqslant 0$$

Suppose that a certain number, N, of events will occur, where N is a Poisson random variable with mean λ. Suppose further that each event that occurs will, independently, be either a type 1 event with probability p or a type 2 event with probability $1-p$. Thus, if N_i is equal to the number of the events that are type i, $i = 1, 2$, then $N = N_1 + N_2$. A useful result is that the random variables N_1 and N_2 are independent Poisson random variables, with respective means

$$E[N_1] = \lambda p \quad E[N_2] = \lambda(1-p)$$

To prove this result, let n and m be nonnegative integers, and consider the joint probability $P\{N_1 = n, N_2 = m\}$. Because $P\{N_1 = n, N_2 = m | N \neq n+m\} = 0$, conditioning on whether $N = n+m$ yields

$$\begin{aligned} P\{N_1 = n, N_2 = m\} &= P\{N_1 = n, N_2 = m | N = n+m\}P\{N = n+m\} \\ &= P\{N_1 = n, N_2 = m | N = n+m\}e^{-\lambda}\frac{\lambda^{n+m}}{(n+m)!} \end{aligned}$$

However, given that $N = n+m$, because each of the $n+m$ events is independently either a type 1 event with probability p or type 2 with probability $1 - p$, it follows that the number of them that are type 1 is a binomial random variable with parameters $n + m$, p. Consequently,

$$\begin{aligned} P\{N_1 = n, N_2 = m\} &= \binom{n+m}{n} p^n (1-p)^m e^{-\lambda} \frac{\lambda^{n+m}}{(n+m)!} \\ &= \frac{(n+m)!}{n!m!} p^n (1-p)^m e^{-\lambda p} e^{-\lambda(1-p)} \frac{\lambda^n \lambda^m}{(n+m)!} \\ &= e^{-\lambda p} \frac{(\lambda p)^n}{n!} e^{-\lambda(1-p)} \frac{(\lambda(1-p))^m}{m!} \end{aligned}$$

Summing over m yields that

$$\begin{aligned} P\{N_1 = n\} &= \sum_m P\{N_1 = n, N_2 = m\} \\ &= e^{-\lambda p} \frac{(\lambda p)^n}{n!} \sum_m e^{-\lambda(1-p)} \frac{(\lambda(1-p))^m}{m!} \\ &= e^{-\lambda p} \frac{(\lambda p)^n}{n!} \end{aligned}$$

Similarly,

$$P\{N_2 = m\} = e^{-\lambda(1-p)} \frac{(\lambda(1-p))^m}{m!}$$

thus verifying that N_1 and N_2 are indeed independent Poisson random variables with respective means λp and $\lambda(1 - p)$.

The preceding result generalizes when each of the Poisson number of events is independently one of the types $1, \ldots, r$, with respective probabilities $p_1, \ldots, p_r$, $\sum_{i=1}^r p_i = 1$. With N_i equal to the number of the events that are type i, $i = 1, \ldots, r$, it is similarly shown that $N_1, \ldots, N_r$ are independent Poisson random variables, with respective means

$$E[N_i] = \lambda p_i, \quad i = 1, \ldots, r$$

Geometric Random Variables

Consider independent trials, each of which is a success with probability p. If X represents the number of the first trial that is a success, then

$$P\{X = n\} = p(1-p)^{n-1}, \quad n \geqslant 1 \tag{2.7}$$

which is easily obtained by noting that in order for the first success to occur on the nth trial, the first $n - 1$ must all be failures and the nth a success. Equation (2.7) now follows because the trials are independent.

A random variable whose probability mass function is given by (2.7) is said to be a geometric random variable with parameter p. The mean of the geometric is obtained as follows:

$$E[X] = \sum_{n=1}^{\infty} np(1-p)^{n-1} = \frac{1}{p}$$

where the above equation made use of the algebraic identity, for $0 < x < 1$,

$$\sum_{n=1}^{\infty} nx^{n-1} = \frac{d}{dx}\left(\sum_{n=0}^{\infty} x^n\right) = \frac{d}{dx}\left(\frac{1}{1-x}\right) = \frac{1}{(1-x)^2}$$

It is also not difficult to show that

$$\text{Var}(X) = \frac{1-p}{p^2}$$

The Negative Binomial Random Variable

If we let X denote the number of trials needed to amass a total of r successes when each trial is independently a success with probability p, then X is said to be a negative binomial, sometimes called a Pascal, random variable with parameters p and r. The probability mass function of such a random variable is given by the following:

$$P\{X = n\} = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \quad n \geqslant r \tag{2.8}$$

To see why Equation (2.8) is valid note that in order for it to take exactly n trials to amass r successes, the first $n-1$ trials must result in exactly $r-1$ successes—and the probability of this is $\binom{n-1}{r-1} p^{r-1}(1-p)^{n-r}$—and then the nth trial must be a success—and the probability of this is p.

If we let $X_i, i = 1, \ldots, r$, denote the number of trials needed after the $(i-1)$st success to obtain the ith success, then it is easy to see that they are independent geometric random variables with common parameter p. Since

$$X = \sum_{i=1}^{r} X_i$$

we see that

$$E[X] = \sum_{i=1}^{r} E[X_i] = \frac{r}{p}$$

$$\text{Var}(X) = \sum_{i=1}^{r} \text{Var}(X_i) = \frac{r(1-p)}{p^2}$$

where the preceding made use of the corresponding results for geometric random variables.

Hypergeometric Random Variables

Consider an urn containing $N + M$ balls, of which N are light colored and M are dark colored. If a sample of size n is randomly chosen [in the sense that each of the $\binom{N+M}{n}$ subsets of size n is equally likely to be chosen] then X, the number of light colored balls selected, has probability mass function

$$P\{X = i\} = \frac{\binom{N}{i}\binom{M}{n-i}}{\binom{N+M}{n}}$$

A random variable X whose probability mass function is given by the preceding equation is called a hypergeometric random variable.

Suppose that the n balls are chosen sequentially. If we let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th selection is light} \\ 0 & \text{otherwise} \end{cases}$$

then

$$X = \sum_{i=1}^{n} X_i \tag{2.9}$$

and so

$$E[X] = \sum_{i=1}^{n} E[X_i] = \frac{nN}{N+M}$$

where the above equation uses the fact that, by symmetry, the ith selection is equally likely to be any of the $N + M$ balls, and so $E[X_i] = P\{X_i = 1\} = N/(N + M)$.

Since the X_i are not independent (why not?), the utilization of the representation (2.9) to compute $\mathrm{Var}(X)$ involves covariance terms. The end product can be shown to yield the result

$$\mathrm{Var}(X) = \frac{nNM}{(N+M)^2}\left(1 - \frac{n-1}{N+M-1}\right)$$

2.9 Continuous Random Variables

In this section we consider certain types of continuous random variables.

Uniformly Distributed Random Variables

A random variable X is said to be uniformly distributed over the interval (a, b), $a < b$, if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

In other words, X is uniformly distributed over (a, b) if it puts all its mass on that interval and it is equally likely to be "near" any point on that interval.

The mean and variance of a uniform (a, b) random variable are obtained as follows:

$$E[X] = \frac{1}{b-a}\int_a^b x dx = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}$$

$$E\left[X^2\right] = \frac{1}{b-a}\int_a^b x^2 dx = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + b^2 + ab}{3}$$

and so

$$\text{Var}(X) = \frac{1}{3}(a^2 + b^2 + ab) - \frac{1}{4}(a^2 + b^2 + 2ab) = \frac{1}{12}(b-a)^2.$$

Thus, for instance, the expected value is, as one might have expected, the midpoint of the interval (a, b).

The distribution function of X is given, for $a < x < b$, by

$$F(x) = P\{X \leqslant x\} = \int_a^x (b-a)^{-1} dx = \frac{x-a}{b-a}$$

Normal Random Variables

A random variable X is said to be normally distributed with mean μ and variance σ^2 if its probability density function is given by

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

The normal density is a bell-shaped curve that is symmetric about μ (see Figure 2.1).

It is not difficult to show that the parameters μ and σ^2 equal the expectation and variance of the normal. That is,

$$E[X] = \mu \qquad \text{and} \qquad \text{Var}(X) = \sigma^2$$

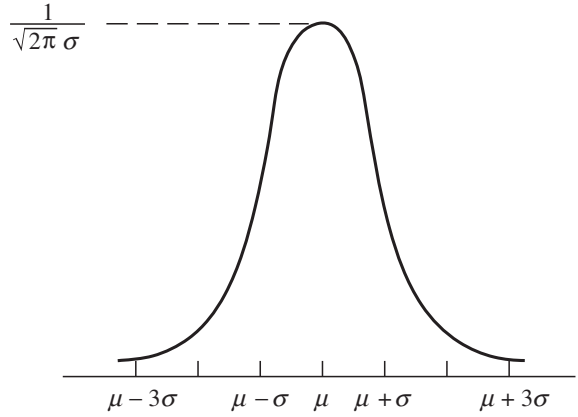

Figure 2.1. The normal density function.

An important fact about normal random variables is that if X is normal with mean μ and variance σ^2, then for any constants a and b, $aX + b$ is normally distributed with mean $a\mu + b$ and variance $a^2\sigma^2$. It follows from this that if X is normal with mean μ and variance σ^2, then

$$Z = \frac{X - \mu}{\sigma}$$

is normal with mean 0 and variance 1. Such a random variable Z is said to have a standard (or unit) normal distribution. Let Φ denote the distribution function of a standard normal random variable; that is,

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

The result that $Z = (X-\mu)/\sigma$ has a standard normal distribution when X is normal with mean μ and variance σ^2 is quite useful because it allows us to evaluate all probabilities concerning X in terms of Φ. For example, the distribution function of X can be expressed as

$$\begin{aligned} F(x) &= P\{X \leqslant x\} \\ &= P\left\{\frac{X - \mu}{\sigma} \leqslant \frac{x - \mu}{\sigma}\right\} \\ &= vP\left\{Z \leqslant \frac{x - \mu}{\sigma}\right\} \\ &= \Phi\left(\frac{x - \mu}{\sigma}\right) \end{aligned}$$

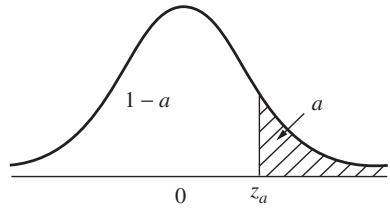

Figure 2.2. $P\{Z > z_a\} = a$.

The value of $\Phi(x)$ can be determined either by looking it up in a table or by writing a computer program to approximate it.

For a in the interval (0, 1), let z_a be such that

$$P\{Z > z_a\} = 1 - \Phi(z_a) = a$$

That is, a standard normal will exceed z_a with probability a (see Figure 2.2). The value of z_a can be obtained from a table of the values of Φ. For example, since

$$\Phi(1.64) = 0.95, \qquad \Phi(1.96) = 0.975, \qquad \Phi(2.33) = 0.99$$

we see that

$$z_{.05} = 1.64, \qquad z_{.025} = 1.96, \qquad z_{.01} = 2.33$$

The wide applicability of normal random variables results fromone of the most important theorems of probability theory—the central limit theorem, which asserts that the sum of a large number of independent random variables has approximately a normal distribution. The simplest form of this remarkable theorem is as follows.

The Central Limit Theorem *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables having finite mean μ and finite variance σ^2. Then*

$$\lim_{n\to\infty} P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right\} = \Phi(x)$$

Exponential Random Variables

A continuous random variable having probability density function

$$f(x) = \lambda e^{-\lambda x}, \quad 0 < x < \infty$$

for some $\lambda > 0$ is said to be an exponential random variable with parameter λ. Its cumulative distribution is given by

$$F(x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}, \quad 0 < x < \infty$$

It is easy to verify that the expected value and variance of such a random variable are as follows:

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

The key property of exponential random variables is that they possess the "memoryless property," where we say that the nonnegative random variable X is memoryless if

$$P\{X > s + t | X > s\} = P\{X > t\} \quad \text{for all } s, t \geqslant 0 \tag{2.10}$$

To understand why the above is called the memoryless property, imagine that X represents the lifetime of some unit, and consider the probability that a unit of age s will survive an additional time t. Since this will occur if the lifetime of the unit exceeds $t + s$ given that it is still alive at time s, we see that

$$P\{\text{additional life of an item of age } s \text{ exceeds } t\} = P\{X > s + t | X > s\}$$

Thus, Equation (2.10) is a statement of fact that the distribution of the remaining life of an item of age s does not depend on s. That is, it is not necessary to remember the age of the unit to know its distribution of remaining life.

Equation (2.10) is equivalent to

$$P\{X > s + t\} = P\{X > s\}P\{X > t\}$$

As the above equation is satisfied whenever X is an exponential random variable—since, in this case, $P\{X > x\} = e^{-\lambda x}$—we see that exponential random variables are memoryless (and indeed it is not difficult to show that they are the only memoryless random variables).

Another useful property of exponential random variables is that they remain exponential when multiplied by a positive constant. To see this suppose that X is exponential with parameter λ, and let c be a positive number. Then

$$P\{cX \leqslant x\} = P\left\{X \leqslant \frac{x}{c}\right\} = 1 - e^{-\lambda x/c}$$

which shows that cX is exponential with parameter λ/c.

Let $X_1, \ldots, X_n$ be independent exponential random variables with respective rates $\lambda_1, \ldots, \lambda_n$. A useful result is that $\min(X_1, \ldots, X_n)$ is exponential with rate $\sum_i \lambda_i$ and is independent of which one of the X_i is the smallest. To verify this, let $M = \min(X_1, \ldots, X_n)$. Then,

$$\begin{aligned} P\{X_j = \min_i X_i | M > t\} &= P\{X_j - t = \min_i (X_i - t) | M > t\} \\ &= P\{X_j - t = \min_i (X_i - t) | X_i > t, i = 1, \ldots, n\} \\ &= P\{X_j = \min_i X_i\} \end{aligned}$$

The final equality follows because, by the lack of memory property of exponential random variables, given that X_i exceeds t, the amount by which it exceeds it is exponential with rate λ_i. Consequently, the conditional distribution of $X_1 - t, \ldots, X_n - t$ given that all the X_i exceed t is the same as the unconditional distribution of $X_1, \ldots, X_n$. Thus, M is independent of which of the X_i is the smallest.

The result that the distribution of M is exponential with rate $\sum_i \lambda_i$ follows from

$$P\{M > t\} = P\{X_i > t, i = 1, \ldots, n\} = \prod_{i=1}^{n} P\{X_i > t\} = e^{-\sum_{i=1}^{n} \lambda_i t}$$

The probability that X_j is the smallest is obtained from

$$\begin{aligned}
P\{X_j = M\} &= \int P\{X_j = M | X_j = t\} \lambda_j e^{-\lambda_j t} dt \\
&= \int P\{X_i > t, i \neq j | X_j = t\} \lambda_j e^{-\lambda_j t} dt \\
&= \int P\{X_i > t, i \neq j\} \lambda_j e^{-\lambda_j t} dt \\
&= \int \left(\prod_{i \neq j} e^{-\lambda_i t} \right) \lambda_j e^{-\lambda_j t} dt \\
&= \lambda_j \int e^{-\sum_i \lambda_i t} dt \\
&= \frac{\lambda_j}{\sum_i \lambda_i}
\end{aligned}$$

The Poisson Process and Gamma Random Variables

Suppose that "events" are occurring at random time points and let $N(t)$ denote the number of events that occur in the time interval $[0, t]$. These events are said to constitute a *Poisson process having rate* λ, $\lambda > 0$, if

(a) $N(0) = 0$.
(b) The numbers of events occurring in disjoint time intervals are independent.
(c) The distribution of the number of events that occur in a given interval depends only on the length of the interval and not on its location.
(d) $\lim_{h \to 0} \frac{P\{N(h)=1\}}{h} = \lambda$.
(e) $\lim_{h \to 0} \frac{P\{N(h) \geqslant 2\}}{h} = 0$.

Thus Condition (a) states that the process begins at time 0. Condition (b), the *independent increment* assumption, states that the number of events by time t [i.e., $N(t)$] is independent of the number of events that occur between t and $t + s$

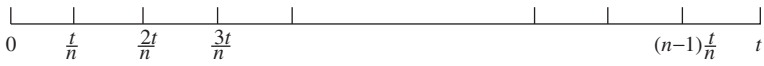

Figure 2.3. The Interval [0, t].

[i.e., $N(t+s) - N(t)$]. Condition (c), the *stationary increment* assumption, states that the probability distribution of $N(t+s) - N(t)$ is the same for all values of t. Conditions (d) and (e) state that in a small interval of length h, the probability of one event occurring is approximately λh, whereas the probability of two or more is approximately 0.

We now argue that these assumptions imply that the number of events occurring in an interval of length t is a Poisson random variable with mean λt. To do so, consider the interval $[0, t]$, and break it up into n nonoverlapping subintervals of length t/n (Figure 2.3). Consider first the number of these subintervals that contain an event. As each subinterval independently [by Condition (b)] contains an event with the same probability [by Condition (c)], which is approximately equal to $\lambda t/n$, it follows that the number of such intervals is a binomial random variable with parameters n and $p \approx \lambda t/n$. Hence, by the argument yielding the convergence of the binomial to the Poisson, we see by letting $n \to \infty$ that the number of such subintervals converges to a Poisson random variable with mean λt. As it can be shown that Condition (e) implies that the probability that any of these subintervals contains two or more events goes to 0 as $n \to \infty$, it follows that $N(t)$, the number of events that occur in $[0, t]$, is a Poisson random variable with mean λt.

For a Poisson process let X_1 denote the time of the first event. Furthermore, for $n > 1$, let X_n denote the elapsed time between the $(n-1)$st and the nth event. The sequence $\{X_n, n = 1, 2, \ldots\}$ is called the *sequence of interarrival times.* For instance, if $X_1 = 5$ and $X_2 = 10$, then the first event of the Poisson process will occur at time 5 and the second at time 15.

We now determine the distribution of the X_n. To do so, we first note that the event $\{X_1 > t\}$ takes place if and only if no events of the Poisson process occur in the interval $[0, t]$; thus

$$P\{X_1 > t\} = P\{N(t) = 0\} = e^{-\lambda t}$$

Hence, X_1 has an exponential distribution with mean $1/\lambda$. To obtain the distribution of X_2, note that

$$\begin{aligned} P\{X_2 > t | X_1 = s\} &= P\{0 \text{ events in } (s, s+t) | X_1 = s\} \\ &= P\{0 \text{ events in } (s, s+t)\} \\ &= e^{-\lambda t} \end{aligned}$$

where the last two equations followed from independent and stationary increments. Therefore, from the foregoing, we conclude that X_2 is also an exponential random variable with mean $1/\lambda$ and, furthermore, that X_2 is independent of X_1. Repeating the same argument yields:

Proposition *The interarrival times $X_1, X_2, \ldots$ are independent and identically distributed exponential random variables with parameter λ.*

Let $S_n = \sum_{i=1}^{n} X_i$ denote the time of the nth event. Since S_n will be less than or equal to t if and only if there have been at least n events by time t, we see that

$$\begin{aligned} P\{S_n \leqslant t\} &= P\{N(t) \geqslant n\} \\ &= \sum_{j=n}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!} \end{aligned}$$

Since the left-hand side is the cumulative distribution function of S_n, we obtain, upon differentiation, that the density function of S_n—call it $f_n(t)$—is given by

$$\begin{aligned} f_n(t) &= \sum_{j=n}^{\infty} j\lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{j!} - \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} \\ &= \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} \\ &= \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \end{aligned}$$

Definition *A random variable having probability density function*

$$f(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}, \quad t > 0$$

is said to be a gamma random variable with parameters (n, λ).

Thus we see that S_n, the time of the nth event of a Poisson process having rate λ, is a gamma random variable with parameters (n, λ). In addition, we obtain from the representation $S_n = \sum_{i=1}^{n} X_i$ and the previous proposition, which stated that these X_i are independent exponentials with rate λ, the following corollary.

Corollary *The sum of n independent exponential random variables, each having parameter λ, is a gamma random variable with parameters (n, λ).*

The Nonhomogeneous Poisson Process

From a modeling point of view the major weakness of the Poisson process is its assumption that events are just as likely to occur in all intervals of equal size. A generalization, which relaxes this assumption, leads to the nonhomogeneous or nonstationary process.

If "events" are occurring randomly in time, and $N(t)$ denotes the number of events that occur by time t, then we say that $\{N(t), t \geqslant 0\}$ constitutes a nonhomogeneous Poisson process with intensity function $\lambda(t), t \geqslant 0$, if

(a) $N(0) = 0$.
(b) The numbers of events that occur in disjoint time intervals are independent.
(c) $\lim_{h\to 0} P\{\text{exactly 1 event between } t \text{ and } t+h\}/h = \lambda(t)$.
(d) $\lim_{h\to 0} P\{2 \text{ or more events between } t \text{ and } t+h\}/h = 0$.

The function $m(t)$ defined by

$$m(t) = \int_0^t \lambda(s)ds, \quad t \geqslant 0$$

is called the mean-value function. The following result can be established.

Proposition *$N(t+s) - N(t)$ is a Poisson random variable with mean $m(t+s) - m(t)$.*

The quantity $\lambda(t)$, called the intensity at time t, indicates how likely it is that an event will occur around the time t. [Note that when $\lambda(t) \equiv \lambda$ the nonhomogeneous reverts to the usual Poisson process.] The following proposition gives a useful way of interpreting a nonhomogeneous Poisson process.

Proposition *Suppose that events are occurring according to a Poisson process having rate λ, and suppose that, independently of anything that came before, an event that occurs at time t is counted with probability $p(t)$. Then the process of counted events constitutes a nonhomogeneous Poisson process with intensity function $\lambda(t) = \lambda p(t)$.*

Proof This proposition is proved by noting that the previously given conditions are all satisfied. Conditions (a), (b), and (d) follow since the corresponding result is true for all (not just the counted) events. Condition (c) follows since

$$\begin{aligned} &P\{1 \text{ counted event between } t \text{ and } t+h\} \\ &\quad = P\{1 \text{ event and it is counted}\} \\ &\qquad + P\{2 \text{ or more events and exactly 1 is counted}\} \\ &\quad \approx \lambda h p(t) \end{aligned}$$

2.10 Conditional Expectation and Conditional Variance

If X and Y are jointly discrete random variables, we define $E[X|Y=y]$, the conditional expectation of X given that $Y = y$, by

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

In other words, the conditional expectation of X, given that $Y = y$, is defined like $E[X]$ as a weighted average of all the possible values of X, but now with the

weight given to the value x being equal to the conditional probability that X equals x given that Y equals y.

Similarly, if X and Y are jointly continuous with joint density function $f(x, y)$, we define the conditional expectation of X, given that $Y = y$, by

$$E[X|Y = y] = \frac{\int xf(x, y)dx}{\int f(x, y)dx}$$

Let $E[X|Y]$ denote that function of the random variable Y whose value at $Y = y$ is $E[X|Y = y]$; and note that $E[X|Y]$ is itself a random variable. The following proposition is quite useful.

Proposition

$$E[E[X|Y]] = E[X] \tag{2.11}$$

If Y is a discrete random variable, then Equation (2.11) states that

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

whereas if Y is continuous with density g, then (2.11) states

$$E[X] = \int E[X|Y = y]\, g(y)dy$$

We now give a proof of the preceding proposition when X and Y are discrete:

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

We can also define the conditional variance of X, given the value of Y, as follows:

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

That is, $\text{Var}(X|Y)$ is a function of Y, which at $Y = y$ is equal to the variance of X given that $Y = y$. By the same reasoning that yields the identity $\text{Var}(X) = E\left[X^2\right] - (E[X])^2$ we have that

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

Taking expectations of both sides of the above equation gives

$$\begin{aligned} E\left[\mathrm{Var}(X|Y)\right] &= E\left[E\left[X^2|Y\right]\right] - E\left[\left(E\left[X|Y\right]\right)^2\right] \\ &= E\left[X^2\right] - E\left[\left(E\left[X|Y\right]\right)^2\right] \end{aligned} \tag{2.12}$$

Also, because $E\left[E\left[X|Y\right]\right] = E\left[X\right]$, we have that

$$\mathrm{Var}\left(E\left[X|Y\right]\right) = E\left[\left(E\left[X|Y\right]\right)^2\right] - \left(E\left[X\right]\right)^2 \tag{2.13}$$

Upon adding Equations (2.12) and (2.13) we obtain the following identity, known as the conditional variance formula.

The Conditional Variance Formula

$$\mathrm{Var}(X) = E\left[\mathrm{Var}(X|Y)\right] + \mathrm{Var}\left(E\left[X|Y\right]\right)$$

Exercises

1.

(a) For any events A and B show that

$$\begin{aligned} A \cup B &= A \cup A^c B \\ B &= AB \cup A^c B \end{aligned}$$

(b) Show that

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

2. Consider an experiment that consists of six horses, numbered 1 through 6, running a race, and suppose that the sample space is given by

$$S = \{ \text{ all orderings of } (1, 2, 3, 4, 5, 6)\}$$

Let A denote the event that the number 1 horse is among the top three finishers, let B denote the event that the number 2 horse comes in second, and let C denote the event that the number 3 horse comes in third.

(a) Describe the event $A \cup B$. How many outcomes are contained in this event?
(b) How many outcomes are contained in the event AB?
(b) How many outcomes are contained in the event ABC?
(c) How many outcomes are contained in the event $A \cup BC$?

3. A couple has two children. What is the probability that both are girls given that the elder is a girl? Assume that all four possibilities are equally likely.

4. The king comes from a family of two children. What is the probability that the other child is his brother?

5. The random variable X takes on one of the values 1, 2, 3, 4 with probabilities

$$P\{X = i\} = ic, \quad i = 1, 2, 3, 4$$

for some value c. Find $P\{2 \leq X \leq 3\}$.

6. The continuous random variable X has a probability density function given by

$$f(x) = cx, \quad 0 < x < 1$$

Find $P\{X > \frac{1}{2}\}$.

7. If X and Y have a joint probability density function specified by

$$f(x, y) = 2e^{-(x+2y)}, \quad 0 < x < \infty, 0 < y < \infty$$

Find $P\{X < Y\}$.

8. Find the expected value of the random variable specified in Exercise 5.

9. Find $E[X]$ for the random variable of Exercise 6.

10. There are 10 different types of coupons and each time one obtains a coupon it is equally likely to be any of the 10 types. Let X denote the number of distinct types contained in a collection of N coupons, and find $E[X]$. [Hint: For $i = 1, \ldots, 10$ let

$$X_t = \begin{cases} 1 & \text{if a type } i \text{ coupon is among the } N \\ 0 & \text{otherwise} \end{cases}$$

and make use of the representation $X = \sum_{i=1}^{10} X_i$.

11. A die having six sides is rolled. If each of the six possible outcomes is equally likely, determine the variance of the number that appears.

12. Suppose that X has probability density function

$$f(x) = ce^x, \quad 0 < x < 1$$

Determine $\text{Var}(X)$.

13. Show that $\text{Var}(aX + b) = a^2\text{Var}(X)$.

14. Suppose that X, the amount of liquid apple contained in a container of commercial apple juice, is a random variable having mean 4 grams.

(a) What can be said about the probability that a given container contains more than 6 grams of liquid apple?
(b) If $\text{Var}(X) = 4(\text{grams})^2$, what can be said about the probability that a given container will contain between 3 and 5 grams of liquid apple?

15. An airplane needs at least half of its engines to safely complete its mission. If each engine independently functions with probability p, for what values of p is a three-engine plane safer than a five-engine plane?

16. For a binomial random variable X with parameters (n, p), show that $P\{X = i\}$ first increases and then decreases, reaching its maximum value when i is the largest integer less than or equal to $(n + 1)p$.

17. If X and Y are independent binomial random variables with respective parameters (n, p) and (m, p), argue, without any calculations, that $X + Y$ is binomial with parameters $(n + m, p)$.

18. Explain why the following random variables all have approximately a Poisson distribution:

(a) The number of misprints in a given chapter of this book.
(b) The number of wrong telephone numbers dialed daily.
(c) The number of customers that enter a given post office on a given day.

19. If X is a Poisson random variable with parameter λ, show that

(a) $E[X] = \lambda$.
(b) $\text{Var}(X) = \lambda$.

20. Let X and Y be independent Poisson random variables with respective parameters λ_1 and λ_2. Use the result of Exercise 17 to heuristically argue that $X + Y$ is Poisson with parameter $\lambda_1 + \lambda_2$. Then give an analytic proof of this. [Hint:

$$P\{X + Y = k\} = \sum_{i=0}^{k} P\{X = i, Y = k - i\} = \sum_{i=0}^{k} P\{X = i\}P\{Y = k - i\}]$$

21. Explain how to make use of the relationship

$$p_{i+1} = \frac{\lambda}{i + 1} p_i$$

to compute efficiently the Poisson probabilities.

22. Find $P\{X > n\}$ when X is a geometric random variable with parameter p.

23. Two players play a certain game until one has won a total of five games. If player A wins each individual game with probability 0.6, what is the probability she will win the match?

24. Consider the hypergeometric model of Section 2.8, and suppose that the white balls are all numbered. For $i = 1, \ldots, N$ let

$$Y_i = \begin{cases} 1 & \text{if white ball numbered } i \text{ is selected} \\ 0 & \text{otherwise} \end{cases}$$

Argue that $X = \sum_{i=1}^{N} Y$, and then use this representation to determine $E[X]$. Verify that this checks with the result given in Section 2.8.

25. The bus will arrive at a time that is uniformly distributed between 8 and 8:30 A.M. If we arrive at 8 A.M., what is the probability that we will wait between 5 and 15 minutes?

26. For a normal random variable with parameters μ and σ^2 show that

(a) $E[X] = \mu$.
(b) $\text{Var}(X) = \sigma^2$.

27. Let X be a binomial random variable with parameters (n, p). Explain why

$$P\left\{\frac{X - np}{\sqrt{np(1-p)}} \le x\right\} \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx$$

when n is large.

28. If X is an exponential random variable with parameter λ, show that

(a) $E[X] = 1/\lambda$.
(b) $\text{Var}(X) = 1/\lambda^2$.

29. Persons A, B, and C are waiting at a bank having two tellers when it opens in the morning. Persons A and B each go to a teller and C waits in line. If the time it takes to serve a customer is an exponential random variable with parameter λ, what is the probability that C is the last to leave the bank? [Hint: No computations are necessary.]

30. Let X and Y be independent exponential random variables with respective rates λ and μ. Is $\max(X, Y)$ an exponential random variable?

31. Consider a Poisson process in which events occur at a rate 0.3 per hour. What is the probability that no events occur between 10 A.M. and 2 P.M.?

32. For a Poisson process with rate λ, find $P\{N(s) = k|N(t) = n\}$ when $s < t$.

33. Repeat Exercise 32 for $s > t$.

34. A random variable X having density function

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, \quad x > 0$$

is said to have *gamma distribution* with parameters $\alpha > 0, \lambda > 0$, where $\Gamma(\alpha)$ is the gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx, \quad \alpha > 0$$

(a) Show that the preceding is a density function. That is, show that it is nonnegative and integrates to 1.
(b) Use integration by parts to show that

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$$

(c) Show that $\Gamma(n) = (n - 1)!, n \geqslant 1$
(d) Find E[X].
(e) Find Var(X).

35. A random variable X having density function

$$f(x) = \frac{x^{a-1}(1 - x)^{b-1}}{B(a, b)}, \quad 0 < x < 1$$

is said to have a *beta distribution* with parameters $a > 0, b > 0$, where $B(a, b)$ is the beta function defined by

$$B(a, b) = \int_0^1 x^{\alpha-1}(1 - x)^{b-1} dx$$

It can be shown that

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

where Γ is the gamma function. Show that $E[X] = \frac{a}{a+b}$.

36. An urn contains four white and six black balls. A random sample of size 4 is chosen. Let X denote the number of white balls in the sample. An additional ball is now selected from the remaining six balls in the urn. Let Y equal 1 if this ball is white and 0 if it is black. Find

(a) $E[Y|X = 2]$.
(b) $E[X|Y = 1]$.
(c) $\text{Var}(Y|X = 0)$.
(d) $\text{Var}(X|Y = 1)$.

37. If X and Y are independent and identically distributed exponential random variables, show that the conditional distribution of X, given that $X + Y = t$, is the uniform distribution on $(0, t)$.

38. Let U be uniform on (0,1). Show that $\min(U, 1 - U)$ is uniform on (0, 1/2), and that $\max(U, 1 - U)$ is uniform on (1/2, 1).

Bibliography

Feller, W., *An Introduction to Probability Theory and Its Applications*, 3rd ed. Wiley, New York, 1968.

Ross, S. M., *A First Course in Probability*, 9th ed. Prentice Hall, New Jersey, 2013.

Ross, S. M., *Introduction to Probability Models*, 10th ed. Academic Press, New York, 2010.