PROBABILITY

Probability theory is a branch of mathematics that deals with randomness and laws of chance. Probability theory is concerned with determining the likelihood of random events, and with characterizing their average or expected behavior. The most fundamental concept of probability theory is the likelihood, or *probability*, of an event. The probability of an event is a number between zero and one, inclusive. Probabilities near one indicate that an event is common or very likely to occur. Probabilities near zero indicate that an event is rare or not very likely to occur. A probability of .5 indicates that it is equally likely for an event to occur or not. Probabilities are often described in percentages (i.e., if an event has a 70% chance of occurring, its probability is .7).

The mathematical foundations of probability theory were developed in the 17th century during a correspondence between Pierre de Fermat and Blaise Pascal about games of chance. This work was expanded in 1713 by Jacques Bernoulli, who derived many early results in combinatorics and games involving *Bernoulli trials* (many repetitions of a procedure with two possible results, such as tossing a coin). In the 18th century, DeMoivre, Laplace, and Gauss developed the normal (or Gaussian) bell-shaped probability distribution to model various physical phenomena. From that time, probability theory was incorporated into many fields at a rapid rate. Today, probability theory has widespread applicability in science and engineering, medicine, social sciences, economics, and actuarial science, as well as many aspects of our everyday lives.

This article begins with an introduction to basic probability principles using games of chance as examples. In games of chance, the results do not depend on the skills of the players but rather on random events such as tossing coins or dice and drawing balls out of a box containing many colored balls. These games are easy to understand and also serve as models for many real-world phenomena. For example, repeated coin tosses model the bits in a binary sequence stored on a disk or transmitted on a digital communications channel.

Modeling more complex phenomena requires the concepts of random variables and probability distributions as well as mathematical techniques from calculus. The remainder of this article covers probability theory at this level. Topics include discrete and continuous random variables, probability distributions, expectation, sums of independent random variables, and limit theorems. There are numerous textbooks devoted to this material. Some representative texts are Refs. 1–6.

Probability theory is the basis for the theory of random, or stochastic, processes. The theory of *stochastic processes* is fundamental to many fields of electrical engineering dealing with signals, including communication theory, signal processing, and control theory. Many textbooks on stochastic processes also include introductory chapters on probability theory geared toward electrical engineers. Some examples are Refs. 7-13. More complex phenomena require advanced treatments of probability and the use of advanced mathematics including linear algebra, real and complex analysis, and measure theory. Some textbooks at this level include Refs. 14–19. Historical details about the development of probability theory can be found in many of the previously referenced texts, especially Ref. 1, and in textbooks on the history of mathematics including Refs. 20 and 21.

BASIC PROBABILITY

This section on basic probability introduces the concepts of random experiments, sample spaces, and events. This framework allows us to describe mathematically our intuitive notions of probability and to develop the concepts of conditional probability, independence, and expectation.

Sample Spaces and Events

The set of all possible results, or *outcomes,* in a game, or *random experiment,* is called the *sample space* and denoted by S. An *event* is a subset of the sample space that contains the desired outcomes. If an experiment has n equally likely outcomes, and f of them are desired, then the probability of the event of interest is f/n. For example, suppose that a random experiment consists of tossing a die. The sample space S contains $n = 6$ equally likely outcomes or *sample points,*

$$S = \{1, 2, 3, 4, 5, 6\}$$

Let A denote the event that an even number appears

$$A = \{2, 4, 6\}$$

and B denote the event that the result is at least three

$$B = \{3, 4, 5, 6\}$$

The event A has 3 sample points; therefore, the probability of A, denoted by $P(A)$, is $\frac{3}{6}$. The event B has four sample points and probability $P(B) = \frac{4}{6}$.

Additional events can be defined in terms of the events A and B using set operations. The *complement* of an event, denoted by A^c consists of all the points in the sample space that are not in the event A. In this example, the complement of A is the set of odd outcomes

$$A^c = \{1, 3, 5\}$$

The *union* of A and B, denoted by $A \cup B$, is the event C, which contains all the sample points in either A or B,

$$C = A \cup B = \{2, 3, 4, 5, 6\}$$

The *intersection* of A and B, denoted by $A \cap B$, is the event D, which contains the sample points that are in both A and B,

$$D = A \cap B = \{4, 6\}$$

The Venn diagrams in Fig. 1 illustrate these relationships.

Two events are *mutually exclusive* if they have no sample points in common (i.e., their intersection is the null set $\emptyset$). In this example, the events D and A^c are mutually exclusive, but A and B are not. The Venn diagram in Fig. 2(a) depicts two mutually exclusive events. A *partition* of an

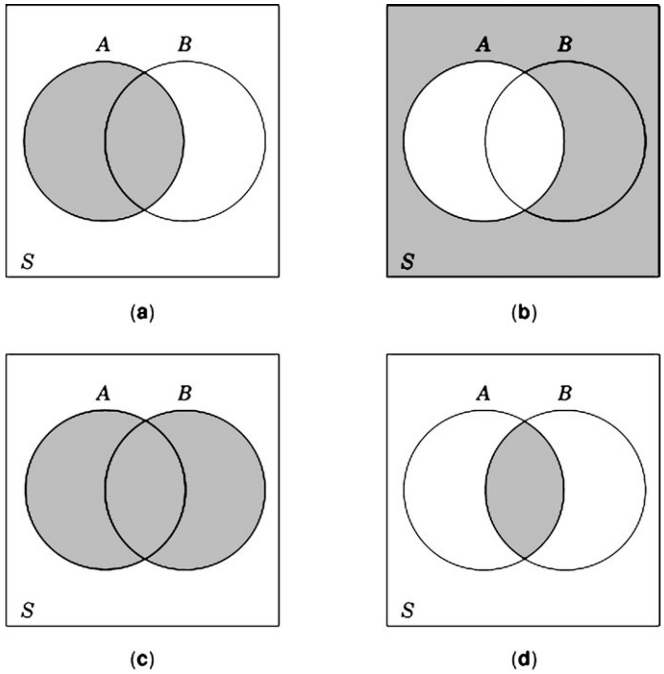

Figure 1. Venn diagrams illustrating set relationships. Shaded regions indicate the event of interest: (a) the event A, (b) the complement of A, (c) the union of A and B, and (d) the intersection of A and B.

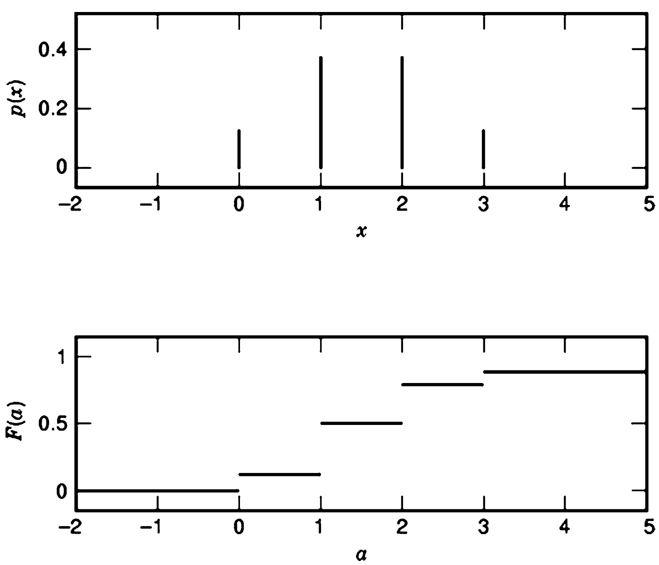

Figure 2. Venn diagrams diagrams illustrating (a) mutually exclusive events and (b) partitioning of A.

event is illustrated in Fig. 2(b). The partition of the event A consists of n smaller events, $A_1, A_2, \ldots, A_n$, which are mutually exclusive and whose union is A (i.e., $A = A_1 \cup A_2 \cup \cdots \cup A_n$). For example, the events $A_1 = \{1\}, A_2 = \{3\}, A_3 = \{5\}$ partition $A = \{1, 3, 5\}$.

Probability Axioms

The basic principles of probability theory can be stated mathematically in terms of sample spaces and events. These are referred to as *probability axioms.*

A1. $P(S) = 1$.
A2. For any event A, $0 \leq P(A) \leq 1$.
A3. For two mutually exclusive events A_1 and A_2, $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.

The first two axioms define probabilities to be numbers between zero and one, with one being the total probability of all the possible outcomes. The third axiom says that if two events cannot occur simultaneously, the probability that either occurs is the sum of their individual probabilities.

Based on these axioms, we can derive additional useful probability rules, such as:

R1. $P(A^c) = 1 - P(A)$.
R2. $P(\emptyset) = 0$.
R3. For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

The first rule says that if an event has a certain portion of the total probability, its complement has the remaining portion. For example $P(A^c) = 1 - P(A) = ½$. The second rule says that the empty set has no probability. The last rule defines the probability of the union of two events that are not mutually exclusive. It is easily verified from the Venn diagrams in Fig. 1. If A and B are not mutually exclusive, then the outcomes common to A and B are counted twice in $P(A) + P(B)$. The probability of $A \cap B$ must be subtracted from $P(A) + P(B)$ to get $P(A \cup B)$. For example $P(A) = \frac{3}{6}$, $P(B) = \frac{4}{6}$, $P(A \cap B) = \frac{2}{6}$, and $P(A \cup B) = 5/6 = \frac{3}{6} + \frac{4}{6} - \frac{2}{6}$.

Counting Techniques

In more complicated games involving multiple coins and dice or colored balls in a box, listing all the possible outcomes can be difficult. For example, suppose four balls are drawn from a box containing seven balls colored red (R), orange (O), yellow (Y), green (G), blue (B), purple (P), and white (W). The balls are drawn one at a time and not put back. The outcome of the four draws is the sequence of colors (e.g., *RWYO* or *GYBW*). We wish to determine the probability that the sequence begins with a red ball followed by a blue ball and contains two additional balls of any color. Listing all the desired outcomes, as well as the possible outcomes in S will be very tedious. Instead, combinatorial analysis can be used to determine the number of outcomes in each set without having to list them. The basic principle is that the total number of outcomes of an experiment consisting of several sequential steps is the product of the number of outcomes at each step. This is known as the *multiplication rule.* In this case, there are seven possible outcomes on the first draw. On the second draw there are only six possibilities because one ball has been removed, on the third draw there are five possibilities, and so on. Thus there are $(7)(6)(5)(4) = 840$ sequences in S. This is an example of a *permutation,* or ordered arrangement, of four objects taken from a group of seven. In general, the number of permutations of r objects taken from a group of n is

$$P_{n,r} = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

For sequences beginning with a red and blue ball, there is only one desired outcome on the first draw (R) and one desired outcome on the second draw (B). On the third draw there are five balls in the box; therefore, there are five possible outcomes, and on the fourth draw there are four possible outcomes. Thus there are a total of $(1)(1)(5)(4) = 20$ desired outcomes. The probability is then $20/840 = \frac{1}{41}$.

In other problems, the order of objects is not important. The number of distinct unordered sets, or *combinations,* of r items taken from a group of n is given by

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

Consider tossing a coin five times. On each toss, there are two possibilities, H or T. The outcome of five tosses is the sequence of heads and tails (e.g., $HTHHT$). We wish to determine the probability of obtaining exactly three heads. The total number of outcomes in the sample space is $(2)(2)(2)(2)(2) = 32$. We are not concerned with the order of the heads and tails; therefore, the number of sequences containing exactly $r = 3$ heads out of the $n = 5$ tosses is $5!/(2!3!) = 120/[(2)(6)] = 10$, and the probability of getting exactly three heads in five tosses is $\frac{10}{32}$.

Conditional Probability

Consider the experiment in which a die is tossed two times. The outcome of the two tosses is a pair of numbers [e.g., (1,2) or (3,6)]. There are six possible outcomes on each toss; therefore, the sample space consists of $(6)(6) = 36$ equally likely outcomes. Suppose that a one is obtained on the first toss and that we wish to determine the probability that the sum of the two tosses will be less than or equal to five. This is an example of a *conditional probability.*

Let $P(B|A)$ denote the conditional probability of the event B given that the event A has occurred. The conditional probability is found from

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad (1)$$

provided $P(A) > 0$.

In this example, we wish to determine the probability that the sum is less than or equal to five given that the first toss is a one; therefore, A is the event that there is a one in the first position,

$$A = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)\}$$

and B is the event that the sum is less than or equal to five

$$B = \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (3,1), (3,2), (4,1)\}$$

Examining the sample points in A, the sum is less than or equal to five for the first four sample points, which are the intersection of A and B,

$$A \cap B = \{(1,1), (1,2), (1,3), (1,4)\}$$

therefore $P(B|A)$ must be $\frac{4}{6}$. We can verify this using the definition in Eq. (1). The required probabilities are $P(A \cap B) = \frac{4}{36}$ and $P(A) = \frac{6}{36}$. As expected, $P(B|A) = (4/36)/(6/36) = \frac{4}{6}$.

We can also reverse the events and find the probability of event A given that event B has occurred from

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (2)$$

provided $P(B) > 0$. In this problem $P(B) = \frac{10}{36}$; therefore, the probability that the first number is a one given that the sum is less than or equal to five is $P(A|B) = (4/36)/(10/36) = \frac{4}{10}$.

The conditional probability formulas in Eqs. (1) and (2) can be rearranged to obtain the *multiplication rules for conditional probability*

$$P(A \cap B) = P(A)P(B|A) \qquad (3)$$

$$P(A \cap B) = P(B)P(A|B) \qquad (4)$$

These relationships are useful in determining probabilities in experiments in which the outcome of a sequence of procedures depends on the previous outcomes. For example, suppose that three boxes contain red and blue balls. Box 1 has two red and two blue balls, Box 2 has one red and two blue balls, and Box 3 has four red and one blue ball. A box is selected at random, and a ball is drawn from the box. What is the probability that a blue ball will be drawn from Box 1? Let A denote selecting Box 1 and B denote drawing a blue ball. Selecting a blue ball from Box 1 is the intersection of events A and B. The probability of selecting Box 1 is $P(A) = \frac{1}{3}$, and the probability of drawing a blue ball given that Box 1 was chosen is $P(B|A) = \frac{2}{4}$; therefore, the probability of selecting a blue ball from Box 1 is $P(A \cap B) = (\frac{1}{3})(\frac{2}{4}) = \frac{1}{6}$.

Suppose that we are also interested in finding the probability of drawing a blue ball from any box. We can find this probability by combining the multiplication rule in Eq. (3) with Axiom 3. Let A_1, A_2, and A_3 denote selecting Boxes 1, 2, and 3, respectively. The events $A_1 \cap B$, $A_2 \cap B$, and $A_3 \cap B$ represent drawing a blue ball from each of the three boxes. They are mutually exclusive events that partition B. From Axiom 3, the probability of drawing a blue ball is the sum of the probabilities of drawing a blue ball from each of the boxes

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

The probability $P(A_1 \cap B) = P(A_1)P(B|A_1)$ has already been found to be $\frac{1}{6}$. Similarly, $P(A_2 \cap B) = P(A_2)P(B|A_2) = (\frac{1}{3})(\frac{2}{3}) = \frac{2}{9}$, and $P(A_3 \cap B) = P(A_3)P(B|A_3) = (\frac{1}{3})(\frac{1}{5}) = \frac{1}{15}$. The total probability of drawing a blue ball is $P(B) = \frac{1}{6} + \frac{2}{9} + \frac{1}{15} = \frac{41}{90}$. This is an example of the *rule of total probability.* In general, if $A_1, A_2, \ldots, A_n$ are mutually exclusive events that partition the sample space S, then for any event B,

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n) \qquad (5)$$

Now suppose that we are told that a blue ball has been drawn and wish to determine the probability that it came from the Box 1 [i.e., $P(A_1|B)$]. Using the definition of conditional probability in Eq. (2), the multiplication rule in Eq. (3), and the rule of total probability in Eq. (5), we have the following result, which is known as *Bayes Rule:*

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(A_1)P(B|A_1)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} \qquad (6)$$

For this example,

$$P(A_1|B) = \frac{(1/3)(2/4)}{(41/90)} = \frac{15}{41}$$

Independence

In some experiments, knowledge about one event tells us nothing about the probability of another event. For example, if a coin is tossed twice, the probability of obtaining a six on the second toss is ⅙, regardless of the outcome of the first toss. The outcomes on the two tosses are said to be *independent.* In general, two events A and B are independent if and only if

$$P(A \cap B) = P(A)P(B) \qquad (7)$$

For this example, let A denote obtaining a one on the first toss and B denote obtaining a six on the second toss. $P(A) = 1/6$, $P(B) = 1/6$, and $P(A \cap B) = \frac{1}{36}$. We see that Eq. (6) is satisfied. Now let C denote the event that the sum is less than or equal to five. From our previous example, $P(C) = \frac{10}{36}$, and $P(A \cap C) = \frac{4}{36}$. In this case, $P(A \cap C) \neq P(A)P(C)$; therefore, A and C are not independent events.

Expectation

Probability theory also deals with expected or long-term average behavior of a random experiment. For example, suppose that a person plays a game in which he or she pays \$1.00 to toss a die. The player wins \$3.00 if a 6 is thrown, \$1.50 if a 5 is thrown, and nothing if a 1, 2, 3, or 4 is thrown. On the average, how much can the player expect to win or lose? Subtracting the cost to play, net winnings are \$2.00 when a 6 is thrown, \$0.50 when a 5 is thrown, and −\$1.00 when a 1, 2, 3, or 4 is thrown. On each toss, the player will win \$2.00 with probability ⅙, \$0.50 with probability ⅙, and −\$1.00 with probability $\frac{4}{6}$. The average winnings are $2(1/6) + 0.5(1/6) - 1(\frac{4}{6}) = -1.5(1/6) = -0.25$. The expected loss per game is 25 cents. Although the player cannot actually lose 25 cents on a given game, this means that if the game is played many times, the player will lose about 25 cents per game on the average.

Summary

These basic concepts form the foundation of probability theory. The examples were based on random experiments in which the sample spaces consisted of a finite number of equally likely events. These experiments serve as models for a many phenomena arising in a variety of applications; however, there are many more phenomena for which these techniques are inadequate. To develop more sophisticated tools for solving more complex problems, we need the concepts of *random variables* and *probability distributions* and mathematical techniques from calculus.

DISCRETE RANDOM VARIABLES AND DISTRIBUTIONS

A random variable is a function that maps the sample points of an experiment onto the real line. For example, suppose that an experiment consists of tossing a coin three times. The sample space contains eight equally likely outcomes

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

If we let X denote the number of heads, then X is a random variable that maps the eight sample points into the numbers 0, 1, 2, and 3. In this case, X is a *discrete random variable* because it can only have one of a discrete set of values.

Probability Mass Function

For discrete random variables, the *probability mass function* (*pmf*) describes how probability is distributed among the possible values of the random variable. The pmf of X is denoted by $p(x)$ and is defined as

$$p(x) = P(X = x) \qquad (8)$$

Probability mass functions have two properties that follow from the probability axioms:

1. $0 \leq p(x) \leq 1$
2. $\Sigma_x \, p(x) = 1$.

Because the pmf is a collection of probabilities, its values must be between zero and one, inclusive. Furthermore, the pmf assigns probability to all the possible values X; therefore, it must sum to one.

In the coin toss problem, the pmf of X is

$$p(0) = P(X = 0) = P(TTT) = 1/8$$
$$p(1) = P(X = 1) = P(TTH) + P(THT) + P(HTT) = 3/8$$
$$p(2) = P(X = 2) = P(THH) + P(HTH) + P(HHT) = 3/8$$
$$p(3) = P(X = 3) = P(HHH) = 1/8$$

It is easy to verify that the properties are satisfied.

Cumulative Distribution Function

The *cumulative distribution function* (*CDF*) also characterizes the probability distribution. The CDF of X is defined for all real numbers a by

$$F(a) = P(X \leq a) = \sum_{x \leq a} p(x) \qquad (9)$$

Some properties of the CDF follow:

1. $\lim_{a \to -\infty} F(a) = 0$.
2. $\lim_{a \to -\infty} F(a) = 1$.
3. $F(a)$ is nondecreasing [i.e., if $a < b$, then $F(a) \leq F(b)$].

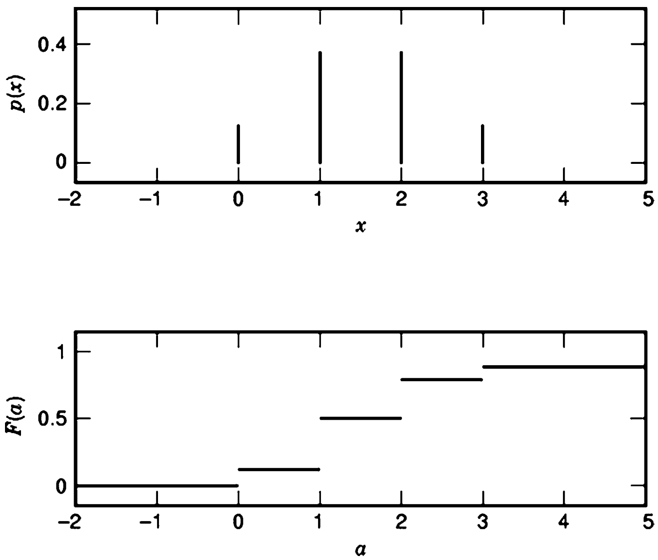

Figure 3. The probability mass function $p(x)$ and cumulative distribution function $F(a)$ of the discrete random variable X.

The CDF for X in the coin toss example is

$$F(a) = \begin{cases} 0 & a < 0 \\ 1/8 & 0 \le a < 1 \\ 1/2 & 1 \le a < 2 \\ 7/8 & 2 \le a < 3 \\ 1 & a \ge 3 \end{cases}$$

For discrete random variables, $F(a)$ has discontinuities or jumps located at the discrete values of the random variable, and the size of a jump is equal to the probability that X is equal to that value. Thus, given the CDF, the pmf can be obtained by subtracting the value of the CDF at the right side of the discontinuity from the value at the left side of the discontinuity. The pmf and CDF for the coin-tossing experiment are shown in Fig. 3.

Probability associated with X can be found from both the pmf and the CDF as follows:

$$P(a < X \le b) = \sum_{a < x \le b} p(x) \tag{10}$$

$$= \sum_{x \le b} p(x) - \sum_{x \le a} p(x)$$

$$= F(b) - F(a) \tag{11}$$

For example, the probability that the number of heads is one or two is $P(0 < X \le 2) = p(1) + p(2) = 3/8 + 3/8 = \frac{6}{8}$, or $P(0 < X \le 2) = F(2) - F(0) = 7/8 - 1/8 = \frac{6}{8}$.

Joint Distribution Functions

The *joint distribution* of two discrete random variables is characterized by the joint pmf and joint CDF. The joint pmf, denoted by $p(x, y)$ assigns probability to all possible joint outcomes

$$p(x, y) = P(X = x, Y = y) \equiv P(X = x \cap Y = y) \tag{12}$$

Similar to the single random variable, or *univariate,* pmf, the joint, or *bivariate,* pmf has the following properties:

Table 1. Joint and Marginal Probability Mass Functions

	$p(x, y)$	y = 0	1	2	$p_X(x)$
	0	.49	.14	.01	.64
x	1	.28	.04	0	.32
	2	.04	0	0	.04
	$p_Y(y)$	.81	.18	.01	1.00

1. $0 \le p(x, y) \le 1$.
2. $\Sigma_{x,y}\, p(x, y) = 1$.

The joint CDF is defined for all real numbers a and b by

$$F(a, b) = P(X \le a, Y \le b) = \sum_{x \le a;\, y \le b} p(x, y) \tag{13}$$

For example, suppose that items produced by an assembly line are tested for defects. Past experience indicates that 70% of the parts have no defects, 20% have minor defects that can be corrected, and 10% have major defects and must be discarded. Suppose that two items are tested. Let X denote the number of items with minor defects and Y denote the number of items with major defects. Then $X \in \{0, 1, 2\}$ and $Y \in \{0, 1, 2\}$. The joint pmf can be found using combinatorial techniques and is summarized in Table 1. Entries for impossible events, such as $(X = 2 \cap Y = 2)$, have zero probability. The probability that X and Y are in some subset A of the possible values can be found from the joint pmf using

$$P(X, Y \in A) = \sum_{x,y \in A} p(x, y) \tag{14}$$

The probability of at least one minor defect and no major defects is, therefore, $P(X \ge 1, Y = 0) = p(1, 0) + p(2, 0) = .28 + .04 = .32$, and the probability of one major defect is $P(Y = 1) = p(0, 1) + p(1, 1) + p(2, 1) = .14 + .04 + 0 = .18$.

Marginal Distributions

The *marginal pmf*s for X alone and Y alone are found from

$$p_X(x) = P(X = x) = \sum_{y} p(x, y) \tag{15}$$

$$p_Y(y = P(Y = y) = \sum_{x} p(x, y) \tag{16}$$

The marginal pmfs are also shown in Table 1.

Conditional Distributions

The *conditional pmf* of X given $Y = y$ is defined as

$$p(x|Y = y) = P(X = x|Y = y)$$

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

$$= \frac{p(x, y)}{p_Y(y)} \quad \text{provided } p_Y(y) > 0 \tag{17}$$

In this example, the pmf of X given $Y = 1$ is

$$p(0|Y=1) = \frac{p(0,1)}{p_Y(1)} = \frac{.14}{.18} = .7778$$
$$p(1|Y=1) = \frac{p(1,1)}{p_Y(1)} = \frac{.04}{.18} = .2222$$
$$p(2|Y=1) = \frac{p(2,1)}{p_Y((1)} = \frac{0}{.18} = 0$$

The conditional pmf of Y given $X = x$ is defined similarly,

$$p(y|X=x) = \frac{p(x,y)}{p_X(x)} \qquad \text{provided } p_X(x) > 0 \qquad (18)$$

Independence

Two discrete random variables are said to be independent if and only if

$$p(x,y) = p_X(x)p_Y(y) \qquad \text{for all } x \text{ and } y \qquad (19)$$

In this example, $p(0, 0) = .49 \neq p_X(0)p_Y(0) = (.64)(.81) = .5184$; therefore, X and Y are not independent.

Transformations

Suppose now that we are interested in the total number of defective items. We can define a new random variable $Z = X + Y$. Then $Z \in \{0, 1, 2\}$. The pmf for Z can be determined as follows

$$p_Z(0) = P(X+Y=0) = p(0,0) = .49$$
$$p_Z(1) = P(X+Y=1) = p(0,1) + P(1,0) = .14 + .28 = .42$$
$$p_Z(2) = P(X+Y=2) = p(0,2) + p(1,1) + p(2,0)$$
$$= 0.1 + .04 + .04 = .09$$

This is an example of a *transformation* of two random variables into a new random variable.

CONTINUOUS RANDOM VARIABLES AND DISTRIBUTIONS

Consider an experiment that consists of monitoring the length of time it takes to serve a customer in a check-out line. The sample space consists of all positive real numbers

$$S = \{y : y > 0\}$$

Let Y denote the length of time it takes to check out a customer. Here Y is a *continuous random variable* because it can have any value on a continuous range, in this case the interval $(0, \infty)$.

Probability Density Function

For continuous random variables, the *probability density function* (*pdf*) denoted by $f(y)$ characterizes the distribution of probability. The probability that Y has a value in the interval $[a, b]$ is found from

$$P(a \le Y \le b) = \int_a^b f(y)\,dy \qquad (20)$$

Thus the probability that Y is in some interval is the area under the pdf over that interval. Note that if $b = a$, then

$$P(a \le Y \le a) = P(Y = a) = \int_a^a f(y)\,dy = 0$$

The probability that Y will assume a particular value is zero; however, this does not mean that it is impossible. For continuous random variables, probability can be assigned only to intervals, not to points. This means that

$$P(a \le Y \le b) = P(a \le Y < b) = P(a < Y \le b)$$
$$= P(a < Y < b) = \int_a^b f(y)\,dy$$

The pdf has the following properties:

1. $f(y) \ge 0$.
2. $\int_{-\infty}^{\infty} f(y)dy = 1$.

The area under the pdf over any interval on the real line is a probability; therefore, the pdf cannot be negative, but it does not necessarily have to be less than one. The pdf integrates to one because the probability that Y is on the real line is one.

A possible pdf for the check-out time Y is

$$f(y) = \begin{cases} e^{-y} & y > 0 \\ 0 & y \le 0 \end{cases}$$

The probability that the check-out time is more than 3 minutes is found from

$$P(Y > 3) = \int_3^{\infty} e^{-y}\,dy = -e^{-y}\Big|_{y=3}^{y=\infty} = e^{-3} = .0498$$

Cumulative Distribution Function

The cumulative distribution function for continuous random variables has the same definition as for discrete random variables and is found from

$$F(a) = P(Y \le a) = \int_{-\infty}^{a} f(y)\,dy \qquad (21)$$

The properties of the CDF are the same as in the discrete case; the probability that Y is in some interval can again be found from

$$P(a < Y \le b) = F(b) - F(a) \qquad (22)$$

For example, the CDF for the check-out time Y is

$$F(a) = \begin{cases} 1 - e^{-a} & a > 0 \\ 0 & a \le 0 \end{cases}$$

The pdf and CDF for check-out times are shown in Fig. 4. Using the CDF, we can find $P(Y > 3) = P(3 < Y < \infty) = F(\infty) - F(3) = 1 - 1 + e^{-3} = .0498$, as expected.

$F(a)$ is a continuous function for continuous random variables, and the pdf $f(y)$ can be obtained from the CDF

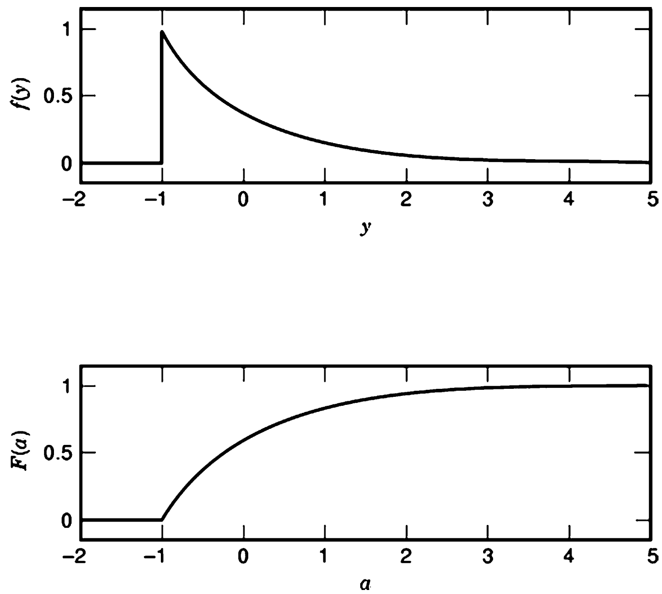

Figure 4. The probability density function $f(y)$ and cumulative distribution function $F(a)$ of the discrete random variable Y.

$F(a)$ by differentiating with respect to a and substituting y for a

$$f(y) = \frac{d}{da}F(a)\Big|_{a=y} \tag{23}$$

Joint Distribution Functions

The joint distribution of two continuous random variables is characterized by the joint pdf and joint CDF. The joint pdf, denoted by $f(x, y)$ assigns probability to regions in the xy plane,

$$P(X, Y \in A) = \iint_A f(x, y)\,dx\,dy \tag{24}$$

The probability that X and Y are within region A is the volume under the pdf over the region A. The joint pdf has the following properties:

1. $f(x, y) \geq 0$.
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dx\,dy = 1$.

The joint CDF for continuous random variables is defined for all real numbers a and b by

$$F(a, b) = P(X \leq a, Y \leq b) = \int_{-\infty}^{b}\int_{-\infty}^{a} f(x, y)\,dx\,dy \tag{25}$$

and the joint pdf can be found from the joint CDF from

$$f(x, y) = \frac{\partial^2}{\partial a \partial b}F(a, b)\Big|_{a=x, b=y} \tag{26}$$

For example, suppose that check-out times for two check-out lines have the following joint pdf:

$$f(x, y) = \begin{cases} 2e^{-(x+2y)} & x > 0, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

The joint CDF is

$$P(X < a, Y < b) = \int_0^b \int_0^a 2e^{-(x+2y)}\,dx dy = (1 - e^{-a})(1 - e^{-2b})$$

The probability that both times are less than 2 minutes is

$$\begin{aligned} P(X < 2, Y < 2) &= \int_0^2 \int_0^2 2e^{-(x+2y)}\,dx\,dy \\ &= (1 - e^{-2})(1 - e^{-4}) = .8488 \end{aligned}$$

Marginal Distributions

Analogous to the discrete case, the marginal pdfs for X alone and Y alone are found from

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy \tag{27}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx \tag{28}$$

The marginal pdfs for our example are

$$f_X(x) = \int_0^{\infty} 2e^{-(x+2y)}dy = e^{-x} \qquad \text{for } x > 0$$

$$f_Y(y) = \int_0^{\infty} 2e^{-(x+2y)}dx = 2e^{-2y} \qquad \text{for } y > 0$$

Conditional Distributions

The conditional pdfs of X given $Y = y$ and Y given $X = x$ are defined as

$$f(x|Y = y) = \frac{f(x, y)}{f_Y(y)} \qquad \text{provided } f_Y(y) > 0 \tag{29}$$

$$f(y|X = x) = \frac{f(x, y)}{f_X(x)} \qquad \text{provided } f_X(x) > 0 \tag{30}$$

In this example, the conditional pdfs are

$$f(x|Y = y) = \frac{2e^{-(x+2y)}}{2e^{-2y}} = e^{-x} \qquad \text{for } x > 0$$

$$f(y|X = x) = \frac{2e^{-(x+2y)}}{e^{-x}} = 2e^{-2y} \qquad \text{for } y > 0$$

In this case, the conditional pdfs are the same as the marginal pdfs.

Independence

Two continuous random variables are said to be independent if and only if

$$f(x, y) = f_X(x) f_Y(y) \qquad \text{for all } x \text{ and } y \tag{31}$$

In this example, $f(x, y) = f_X(x) f_Y(y) = 2e^{-(x+2y)}$ for $x > 0$ and $y > 0$; therefore, X and Y are independent.

Transformations

Suppose that it costs the store \$5.00 to process each customer plus \$2.00 for each minute spent at check-out. Let Z be the cost to check out through the second line (i.e., $Z = 5 + 2X$). This is an example of a transformation of a continuous random variable. In general, if $Z = g(X)$ and $g(X)$ is an invertible function, then $X = g^{-1}(Z)$ and the pdf of Z is given by

$$f_Z(z) = f_X(g^{-1}(z)) \left| \frac{d}{dz} g^{-1}(z) \right| \tag{32}$$

In our example $Z = g(X) = 5 + 2X$; therefore, $X = g^{-1}(Z) = (Z - 5)/2$ and

$$f_Z(z) = 2e^{-2(z-5)/2} \left| \frac{1}{2} \right| = e^{-(z-5)} \qquad \text{for } z > 5$$

This result can be generalized to transformations of multiple random variables.

EXPECTATION

Expected Values

The statistical average or *expected value* of a random variable is denoted by $E[X]$ or μ. For discrete random variables, it is defined as

$$\mu \equiv E[X] = \sum_x x p(x) \tag{33}$$

The expected value of X is a weighted average of the possible values of the random variable, with the weighting determined by the probability of each value. In the coin tossing example, $E[X] = 0(\frac{1}{8}) + 1(\frac{3}{8}) + 2(\frac{3}{8}) + 3(\frac{1}{8}) = 1.5$.

For continuous random variables, the expected value of X is defined as

$$\mu \equiv E[X] = \int_{-\infty}^{\infty} x f(x)\, dx \tag{34}$$

In the check-out time example,

$$E[Y] = \int_0^{\infty} y e^{-y}\, dy = 1$$

The expected value of X is also called the *mean* and has the interpretation as the center of mass of the pmf or pdf.

The expected value of a function of X, say $g(X)$, is given by

$$E[g(X)] = \sum_x g(x) p(x) \qquad \text{(discrete)} \tag{35}$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \qquad \text{(continuous)} \tag{36}$$

In particular, if $g(X) = X^k$, the expected value is known as the *kth moment* of X. Thus the mean is the first moment. If $g(X) = (X - \mu)^k$, this is called the *kth central moment* of X. When $n = 2$, the second central moment is called the *variance* and denoted by σ^2. The variance is related to the first and second moments by

$$\sigma^2 \equiv E[(X - \mu)^2] = E[X^2] - \mu^2 \tag{37}$$

The square root of the variance (σ) is called the *standard deviation.* The variance and standard deviation are measures of the spread of the distribution from the mean. A distribution concentrated close to the mean will have a small standard deviation, and a widely spread distribution will have a large standard deviation. *Chebychev's Inequality,* which is discussed at the end of the article, provides the general rule of thumb that most of the probability is found within two to three standard deviations from the mean.

In the coin tossing example, $E[X^2] = 0(\frac{1}{8}) + 1(\frac{3}{8}) + 4(\frac{3}{8}) + 9(\frac{1}{8}) = 3$, $\sigma^2 = 3 - (1.5)^2 = .75$ and $\sigma = \sqrt{.75} = .8660$. In the check-out time example, $\sigma^2 = 2 - (1)^2 = 1$ and $\sigma = \sqrt{1} = 1$.

Moment Generating Function

The *moment generating function* (*MGF*) of the random variable X is defined as

$$M(t) = E[e^{Xt}] \tag{38}$$

The kth moment of X can be obtained from $M(t)$ by differentiating k times with respect to t and setting $t = 0$,

$$E[X^k] = \frac{d^k}{dt^k} M(t) \bigg|_{t=0} \tag{39}$$

In finding the first two moments, we use the notation

$$M'(t) \equiv \frac{d}{dt} M(t) \tag{40}$$

$$M''(t) \equiv \frac{d^2}{dt^2} M(t) \tag{41}$$

therefore, $E[X] = M'_X(0)$ and $E[X^2] = M''_X(0)$.

In the coin-tossing example, $M_X(t) = \frac{1}{8} + (\frac{3}{8})e^t + (\frac{3}{8})e^{2t} + (\frac{1}{8})e^{3t}$, $M'_X(t) = (\frac{3}{8})e^t + (\frac{6}{8})e^{2t} + (\frac{3}{8})e^{3t}$, and $M''_X(t) = (\frac{3}{8})e^t + (\frac{12}{8})e^{2t} + (\frac{9}{8})e^{3t}$. Therefore, $M'_X(0) = 1.5 = E[X]$ and $M''_X(0) = 3 = E[X^2]$. In the check-out time example, $M_Y(t) = (1 - t)^{-1}$, $M'_Y(t) = (1 - t)^{-2}$, and $M''_Y(t) = 2(1 - t)^{-3}$. Therefore, $M'_Y(0) = 1 = E[Y]$ and $M''_Y(0) = 2 = E[Y^2]$.

Joint Expectation

Expected values for jointly distributed random variables are defined as

$$E[g(X, Y)] = \sum_{x,y} g(x, y) p(x, y) \qquad \text{(discrete)} \tag{42}$$

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy \qquad \text{(continuous)} \tag{43}$$

Joint moments are obtained when $g(X, Y) = X^k Y^m$, and joint central moments are obtained when $g(X, Y) = (X - \mu_X)^k (Y - \mu_Y)^m$. When $k = m = 1$, the joint central moment is called the *covariance* of X and Y. It is related to the joint and marginal moments by

$$\mathrm{COV}(X, Y) \equiv E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y \tag{44}$$

Properties

Let $Z = aX + bY + c$, then

$$\mu_Z = a\mu_X + b\mu_Y + c \tag{45}$$

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\ \mathrm{COV}(X, Y) \tag{46}$$

If the random variables X and Y are independent, then

$$E[g_1(X)g_2(Y)] = E[g_1(X)]E[g_2(Y)] \tag{47}$$

This means that $E[X, Y] = E[X]E[Y] = \mu_X\mu_Y$ and COV(X, Y) = 0. (The converse is not necessarily true, i.e., if COV(X, Y) = 0, X and Y are not necessarily independent.) Then the variance of Z reduces to

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 \tag{48}$$

Another consequence of Eq. (47) is that the moment generating function of Z is

$$\begin{aligned} M_Z(t) &= E[e^{(aX+bY+c)t}] = e^{ct}E[e^{aXt}]E[e^{bYt}] \\ &= e^{ct}M_X(at)M_Y(bt) \end{aligned} \tag{49}$$

SUMS OF INDEPENDENT RANDOM VARIABLES

A random sample of size n is a sequence of random variables $X_1, X_2, \ldots, X_n$, which are *independent and identically distributed* (*i.i.d.*). Their *multivariate* joint pmf or pdf has the form

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2)\cdots p(x_n) \qquad \text{(discrete)} \tag{50}$$

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_1)\cdots f(x_n) \qquad \text{(continuous)} \tag{51}$$

As a consequence, COV(X_iX_j) = 0 for $i \neq j$.

We are often interested in sums of i.i.d. random variables. Let $Z = X_1 + X_2 + \cdots + X_n$. We have the following results:

$$\mu_Z = \mu_X + \mu_X + \cdots + \mu_X = n\mu_X \tag{52}$$

$$\sigma_Z^2 = \sigma_X^2 + \sigma_X^2 + \cdots + \sigma_X^2 = n\sigma_X^2 \tag{53}$$

$$M_Z(t) = M_X(t)M_X(t)\cdots M_X(t) = M_X(t)^n \tag{54}$$

Another case of interest is $\bar{X} = (1/n)(X_1 + X_2 + \cdots + X_n)$. This is known as the *sample mean.* The mean, variance, and MGF of $\bar{X}$ are

$$\mu_{\overline{X}} = \frac{1}{n}(\mu_X + \mu_X + \cdots + \mu_X) = \mu_X \tag{55}$$

$$\sigma_{\overline{X}}^2 = \frac{1}{n^2}(\sigma_X^2 + \sigma_X^2 + \cdots + \sigma_X^2) = \frac{\sigma_X^2}{n} \tag{56}$$

$$M_{\overline{X}}(t) = M_X(t/n)M_X(t/n)\cdots M_X(t/n) = M_X(t/n)^n \tag{57}$$

SPECIAL DISCRETE DISTRIBUTIONS

Certain discrete random variables arise frequently in modeling physical phenomena. Some special discrete random variables are described below and summarized in Table 2.

Bernoulli (p)

A Bernoulli random variable X is a discrete random variable which has two possible outcomes, 1 and 0, with probabilities p and $1 - p$. An experiment whose outcome is a Bernoulli random variable is called a *Bernoulli trial.* The Bernoulli random variable is named after Swiss mathematician Jacques (Jakob, James) Bernoulli, who studied games involving many repetitions of a procedure with two possible outcomes. He derived many early results in combinatorics including the formulas for permutations and combinations as well as the binomial expansion. His work in this area was published in 1713, eight years after his death.

The Bernoulli random variable models things like data bits, operational status of equipment (on or off), test results (pass or fail), quality of manufactured items (defective or good), and so on. The result $X = 1$ is usually called a success, and the result $X = 0$ is called a failure. The pmf of X is

$$p(x) = \begin{cases} 1 - p & x = 0 \\ p & x = 1 \end{cases} \tag{58}$$

The first two moments of X are

$$E[X] = 0(1 - p) + 1(p) = p$$
$$E[X^2] = 0(1 - p) + 1(p) = p$$

Therefore, the mean and variance are

$$\mu = p$$
$$\sigma^2 = p(1 - p)$$

The MGF of X is

$$M_X(t) = e^{0t}(1 - p) + e^{1t}p = 1 - p + pe^t$$

Taking derivatives

$$M_X'(t) = pe^t$$

$$M_X''(t) = pe^t$$

therefore, $E[X] = M_X'(0) = p$ and $E[X_2] = M_X''(0) = p$, as expected. The Bernoulli random variable is the basic building block for the Binomial, Geometric, and Negative Binomial random variables, which characterize different observations of repeated Bernoulli trials.

Binomial (n, p)

A Binomial random variable X is the number of successes obtained in n i.i.d. Bernoulli trials. The possible values for X are 0, 1, …, n. The probability that $X = 0$ is the probability of n failures, which is $(1 - p)^n$, because the trials are independent and the probabilities multiply. The probability that $X = 1$ is the probability of $n - 1$ failures and 1 success, multiplied by the number of combinations of $r = 1$ successes out of n trials,

$$p(1) = \binom{n}{1}p(1 - p)^{n-1}$$

In general, the probability that $X = x$ is the probability of x successes and $n - x$ failures multiplied by the number of

Table 2. Discrete Random Variables

Type	Parameters	$p(x)$	x	μ	σ^2	$M(t)$
Bernoulli	p	$p^x(1-p)^{1\ x}$	0, 1	p	$p(1-{}^{\ast})$	$1-p+pe^t$
Binomial	n, p	$\binom{n}{x}p^x(1-p)^{n\ x}$	0, 1, ..., n	np	$np(1-p)$	$(1-p-pe^t)^n$
Geometric	p	$p(1-p)^{x\ 1}$	1, 2, ...	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$	$\frac{pe^t}{1\ (1\ p)e^t}$
Negative Binomial	k, p	$\binom{x\ \ 1}{k-1}p^k(1-p)^{x\ k}$	k, k + 1, ...	$\frac{k}{p}$	$\frac{k(1-p)}{p^2}$	$\left[\frac{pe^t}{1-(1-p)e^t}\right]^k$
Poisson	μ	$\frac{e\ ^{\mu}(\mu)^x}{x!}$	0, 1, 2, ...	μ	μ	$e^{\mu(e^t\ 1)}$

combinations of x successes in n trials. Therefore,

$$p(x) = \binom{n}{x} p^x(1-p)^{n-x} \qquad x = 0, 1, \ldots, n \qquad (59)$$

The factor

$$\binom{n}{x}$$

is called a *binomial coefficient* because it appears in the *binomial expansion* of the sum of two numbers raised to the power n:

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} \qquad (60)$$

The binomial expansion can be used to show that

$$\sum_{x=0}^{n} p(x) = \sum_{x=0}^{n} \binom{n}{x} p^x(1-p)^{n-x} = (p+1-p)^n = 1$$

It is often convenient to express X as the sum of n i.i.d. Bernoulli random variables, $X_1, \ldots, X_n$. Therefore when $n=1$, a Binomial (1, p) random variable is the same as a Bernoulli (p) random variable. Using the properties of sums of independent random variables, the mean and variance of X are $\mu = np$ and $\sigma^2 = np(1-p)$, and the MGF of X is $M(t) = (1-p+pe^t)^n$.

Geometric (p)

A Geometric random variable X is the number of Bernoulli trials until the first success is obtained. The possible values for X are 1, 2, ... The probability that $X=1$ is the probability that a success is obtained on the first trial, which is p. The probability that $X=2$ is the probability of a failure on the first trial and a success on the second trial, which is $(1-p)p$. In general, the probability that $X=x$ is the probability of $x-1$ failures and a success on the xth trial. Therefore,

$$p(x) = p(1-p)^{x-1} \qquad x = 1, 2, \ldots \qquad (61)$$

The mean and variance of X are $\mu = 1/p$ and $\sigma^2 = 1/p^2$, and the MGF of X is $M(t) = pe^t/[1-(1-p)e^t]$.

These can be derived using the *geometric series*

$$\sum_{x=0}^{\infty} a^x = \frac{1}{1-a} \qquad |a| < 1 \qquad (62)$$

Negative Binomial (k, p)

A Negative Binomial random variable X is the number of Bernoulli trials until the kth success is obtained. The possible values for X are $k, k+1, \ldots$. The probability that $X=k$ is the probability that k successes are obtained on the first k trials, which is p^k. The probability that $X=k+1$ is the probability of $k-1$ successes and one failure on the first k trials, and a success on the $(k+1)$th trial, multiplied by the number of combinations of $k-1$ successes in k trials. In general, the probability that $X=x$ is the probability of $k-1$ successes and $x-k$ failures on the first $x-1$ trials and a success on the xth trial, multiplied by the number of combinations of $k-1$ successes in $x-1$ trials. Therefore,

$$p(x) = \binom{x-1}{k-1} p^k(1-p)^{x-k} \qquad x = k, k+1, \ldots \qquad (63)$$

X can be expressed as the sum of n i.i.d. Geometric random variables, $X_1, \ldots, X_n$. Using the properties of sums of independent random variables, the mean and variance of X are $\mu = k/p$ and $\sigma^2 = k/p^2$, and the MGF of X is $M(t) = \{pe^t/[1-(1-p)e^t]\}^k$. The Negative Binomial distribution gets its name because proving that the distribution sums to one requires use of the binomial expansion of $[1-(1-p)]$ raised to a negative power $(-k)$. It is also called the Pascal distribution after French mathematician Blaise Pascal.

Poisson (μ)

A Poisson random variable X is the number of successes observed in an interval of length $\mu = \lambda t$, where λ is the average rate of successes and t is an interval of observation. The interval may correspond to time, length, etc. The possible values for X are 0, 1, ..., and the pdf is

$$p(x) = \frac{e^{-\mu}(\mu)^x}{x!} \qquad x = 0, 1, 2, \ldots \qquad (64)$$

The mean and variance of X are $\mu = \mu$ and $\sigma^2 = \mu$, and the MGF of X is $M(t) = e^{\mu(e^t-1)}$. These can be derived using the series expansion of the exponential function,

$$\sum_{x=0}^{\infty} \frac{a^x}{x!} = e^a \qquad (65)$$

The Poisson distribution gets its name from French mathematician Simeon-Denis Poisson who introduced it in 1837 as a limiting form of a binomial distribution when n becomes large and p becomes small while np remains constant.

SPECIAL CONTINUOUS DISTRIBUTIONS

Some special continuous random variables are described below and summarized in Table 3.

Uniform (a, b)

A uniform random variable X is a continuous random variable whose pdf is constant, or uniform, over the interval [a, b]. To ensure that the area under the pdf is one, the height of the pdf must be the inverse of the length of the interval $1/(b-a)$

$$f(x) = \frac{1}{b-a} \qquad a \le x \le b \tag{66}$$

The mean and variance are $\mu = (a+b)/2$ and $\sigma^2 = (b-a)^2/12$, and the MGF of X is $M(t) = (e^{tb} - e^{ta})/[t(b-a)]$. At $t = 0$, the MGF has the form $\frac{0}{0}$; however, it is easy to verify that the limit as $t \to 0$ exists and is equal to one. The derivatives of the MGF also have the form $\frac{0}{0}$, but the limits as $t \to 0$ exist and are equal to the moments.

Normal (μ, σ^2)

A normal random variable X has the pdf

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty \tag{67}$$

The mean and variance are μ and σ^2, and the MGF is $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$. The *standard normal* distribution has $\mu = 0$ and $\sigma^2 = 1$. This distribution was first introduced by French–English mathematician Abraham DeMoivre in 1733 as an approximation to the binomial distribution. He called it the exponential bell-shaped curve. The normal distribution is also called the Gaussian distribution after German mathematician and scientist Karl Friedrich Gauss, who used it to model errors in scientific experiments in 1809. It was referred to as the normal distribution by British statistician Karl Pearson in the late 19th century, who observed that it was "normal" for data sets to have this distribution. These observations are consequences of the Central Limit Theorem, which is discussed at the end of the article. It states that the distribution of a sum of a large number of independent random variables is approximately normal. Because of this result, the normal distribution models many phenomena.

Gamma (α, β)

The gamma random variable has pdf

$$f(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \qquad x \ge 0 \tag{68}$$

where $\Gamma(\alpha)$ is the gamma function, defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx \tag{69}$$

For $\alpha > 1$, the gamma function has the property

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

When α is a positive integer (i.e., $\alpha = n$), $\Gamma(n) = (n-1)!$. The mean and variance are of X are $\mu = \alpha\beta$ and $\sigma^2 = \alpha\beta^2$, and the MGF is $M(t) = (1-\beta t)^{-\alpha}$. Some special cases of the Gamma distribution include the Exponential, Erlang, and Chi-Square distributions.

Exponential (λ)

The Exponential distribution is obtained when $\alpha = 1$ and $\beta = 1/\lambda$. The pdf is

$$f(x) = \lambda e^{-\lambda x} \qquad x \ge 0 \tag{70}$$

The mean and variance are $1/\lambda$ and $1/\lambda^2$, and the MGF is $M(t) = [1-(t/\lambda)]^{-1}$. The exponential distribution models the time between successes when the number of successes has a Poisson distribution with $\mu = \lambda t$. The parameter λ is the average rate of success in both distributions.

Erlang (n, λ)

The Erlang distribution is obtained when $\alpha = n$ and $\beta = 1/\lambda$. The pdf is

$$f(x) = \frac{(\lambda x)^{n-1} \lambda e^{-\lambda x}}{(n-1)!} \qquad x \ge 0 \tag{71}$$

The mean and variance are n/λ and n/λ^2, and the MGF is $M(t) = [1-(t/\lambda)]^{-n}$. The Erlang distribution models the time until n successes occur when the number of successes has a Poisson distribution with $\mu = \lambda t$. An Erlang random variable is the sum of n i.i.d. Exponential random variables. It is named after Danish engineer and mathematician A. K. Erlang, who studied call traffic in telephone systems.

Chi-Square (n)

The Chi-Square distribution is obtained when $\alpha = n/2$ and $\beta = 2$. The pdf is

$$f(x) = \frac{x^{\frac{n}{2}-1} e^{-x/2}}{2^{n/2}\Gamma(n/2)} \qquad x \ge 0 \tag{72}$$

The mean and variance are n and $2n$, and the MGF is $M(t) = (1-2t)^{-n/2}$. The Chi-Square distribution is obtained from the sum of the squares of n standard normal random variables.

Rayleigh (σ_r^2)

The pdf of the Rayleigh distribution is

$$f(x) = \frac{x e^{-x^2/2\sigma_r^2}}{\sigma_r^2} \qquad x \ge 0 \tag{73}$$

The mean and variance are $\mu = \sigma_r\sqrt{\pi/2}$ and $\sigma^2 = \sigma^2{}_r(2-\pi/2)$, and the MGF is

$$M(t) = 1 + \sqrt{\pi}\sigma_r t e^{\sigma_r^2 t^2/2}\{1/\sqrt{2} + \mathrm{erf}(\sigma t/\sqrt{2})\} \tag{74}$$

Table 3. Continuous Random Variables

Type	Parameters	$f(x)$	x	μ	σ^2	$M(t)$
Uniform	a, b	$\frac{1}{b-a}$	$[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Normal	μ, σ^2	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$	$(-\infty, \infty)$	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$
Gamma	α, β	$\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}$	$[0, \infty)$	$\alpha\beta$	$\alpha\beta^2$	$(1-\beta t)^{-\alpha}$
Exponential	λ	$\lambda e^{-\lambda x}$	$[0, \infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$(1-t/\lambda)^{-1}$
Erlang	n, λ	$\frac{(\lambda x)^{n-1}\lambda e^{-\lambda x}}{(n-1)!}$	$[0, \infty)$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$	$(1-t/\lambda)^{-n}$
Chi-Square	n	$\frac{x^{(n/2)-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}$	$[0, \infty)$	n	$2n$	$(1-2t)^{-n/2}$
Rayleigh	σ_r^2	$\frac{xe^{-x^2/2\sigma_r^2}}{\sigma_r^2}$	$[0, \infty)$	$\sigma_r\sqrt{\frac{\pi}{2}}$	$\sigma_r^2\left(2-\frac{\pi}{2}\right)$	Eq. (74)

where the *error function* is defined as

$$\text{erf}(z) = \sqrt{\frac{\pi}{2}}\int_0^z e^{v^2}\,dv \tag{75}$$

The Rayleigh distribution is obtained from the square root of sum of the squares of two normal $(0, \sigma_r^2)$ random variables.

LIMIT THEOREMS

Two of the most important theorems in probability theory are the *Law of Large Numbers* and the *Central Limit Theorem.* Another important result is *Chebychev's Inequality,* upon which the Weak Law of Large Numbers is based. Here we state these theorems without giving proofs.

Chebychev's Inequality

If X is a random variable with mean μ and variance σ^2, then for any $k > 0$,

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{76}$$

For example, let $k = 2$. This inequality says that the probability that X has a value more than two standard deviations from its mean is less than .25. The probability that X is more than three standard deviations from its mean is less than .10. This theorem has many important theoretical implications, but it also has practical ones. A good rule of thumb for both discrete and continuous random variables is that most of the probability is within a few standard deviations of the mean.

Weak Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Then for any $\epsilon > 0$

$$P(|\overline{X}-\mu| \geq \epsilon) \to 0 \text{ as } n \to \infty \tag{77}$$

From Eqs. (52) and (53), the mean and variance of $\bar{X}$ are μ and σ^2/n. Using Chebychev's Inequality with $\epsilon = k\sigma/n$ and $k^2 = \epsilon^2 n/\sigma^2$,

$$P(|\overline{X}-\mu| \geq \epsilon) = P\left(|\overline{X}-\mu| \geq k\frac{\sigma}{\sqrt{n}}\right) \leq \frac{\sigma^2}{\epsilon^2 n}$$

As $n \to \infty$, this probability goes to zero. The Law of Large Numbers ensures that the average of a set of i.i.d. random variables converges to their mean as the number of samples increases.

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. The distribution of

$$Z = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal distribution as $n \to \infty$.

Applying the properties of sums of independent random variables, the mean of Z is zero and its variance is one. The Central Limit Theorem says that the distribution of Z approaches the standard normal distribution for large n, regardless of the distribution of the sample. This theorem proves what is often observed in practice, namely that the sum of a large number of i.i.d. random variables has a distribution which is approximately normal.

SUMMARY

The concepts of random experiments, sample spaces, and events provide the framework to mathematically describe the principles of probability and to develop the concepts of conditional probability, independence, and expectation. Random variables and probability distributions provide the additional tools necessary to analyze a wide range of random phenomena. In this article, we have provided an introduction to discrete and continuous random variables, probability distributions, and expectation; developed properties for sums of independent random variables; and introduced two important probability theorems, the Law of Large Numbers and the Central Limit Theorem. We also summarized some important discrete and continuous probability distributions.

BIBLIOGRAPHY

1. S. Ross, *A First Course in Probability*, 6th ed., Upper Saddle River, NJ: Prentice-Hall, 2001.
2. J. Pitman, *Probability*, New York: Springer-Verlag, 1999.
3. J. J. Higgins and S. Keller-McNulty, *Concepts in Probability and Stochastic Modeling*, Belmont, CA: Duxbury Press, 1995.
4. R. E. Walpole, R. H. Myers, and S. L. Myers, *Probability and Statistical for Engineers and Scientists*, 7th ed., Upper Saddle River, NJ: Prentice Hall, 2002.
5. P. L. Meyer, *Introductory Probability and Statistical Applications*, 2nd ed., Reading, MA: Addison-Wesley, 1970.
6. R. V. Hogg and A. T. Craig, *Introduction to Mathematical Statistics*, 6th ed., New York: Macmillan, 2004.
7. W. B. Davenport, Jr. and W. L. Root *An Introduction to the Theory of Random Signals and Noise*, New York: IEEE Press, 1987. Originally published by McGraw-Hill in 1958.
8. E. Cinlar, *Introduction to Stochastic Processes*, Englewood, NJ: Prentice-Hall, 1974.
9. H. Stark and J. W. Woods *Probability, Random Processes, and Estimation Theory for Engineers*, 3rd ed., Englewood Cliffs, NJ: Prentice-Hall, 2001.
10. A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed., New York: McGraw-Hill, 2002.
11. C. W. Helstrom, *Probability and Stochastic Processes for Engineers*, 2nd ed., New York: Macmillan, 1991.
12. A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, 2nd ed., Reading, MA: Addison-Wesley, 1994.
13. P. Z. Peebles Jr., *Probability, Random Variables, and Random Signal Principles*, 4th ed., New York: McGraw-Hill, 2000.
14. M. Loeve, *Probability Theory*, 3rd ed., Princeton, NJ: Van Nostrand, 1963.
15. W. Feller, *An Introduction to Probability Theory and Its Application, Vol. I*, 3rd ed., New York: Wiley, 1968.
16. W. Feller, *An Introduction to Probability Theory and Its Application, Vol. II*, 2nd ed., New York: Wiley, 1971.
17. L. Breiman, *Probability*, Philadelphia: SIAM, 1992. Originally published by Addison-Wesley in 1968.
18. P. Billingsley, *Probability and Measure*, 2nd ed., New York: Wiley, 1986.
19. R. M. Gray, *Probability, Random Processes, and Ergodic Properties*, New York: Springer-Verlag, 1988.
20. L. Motz and J. H. Weaver, *The Story of Mathematics*, New York: Plenum, 1993.
21. C. Boyer and U. C. Merzbach, *A History of Mathematics*, rev., 2nd ed., New York: Wiley, 1991.

KRISTINE L. BELL
George Mason University,
Fairfax, VA