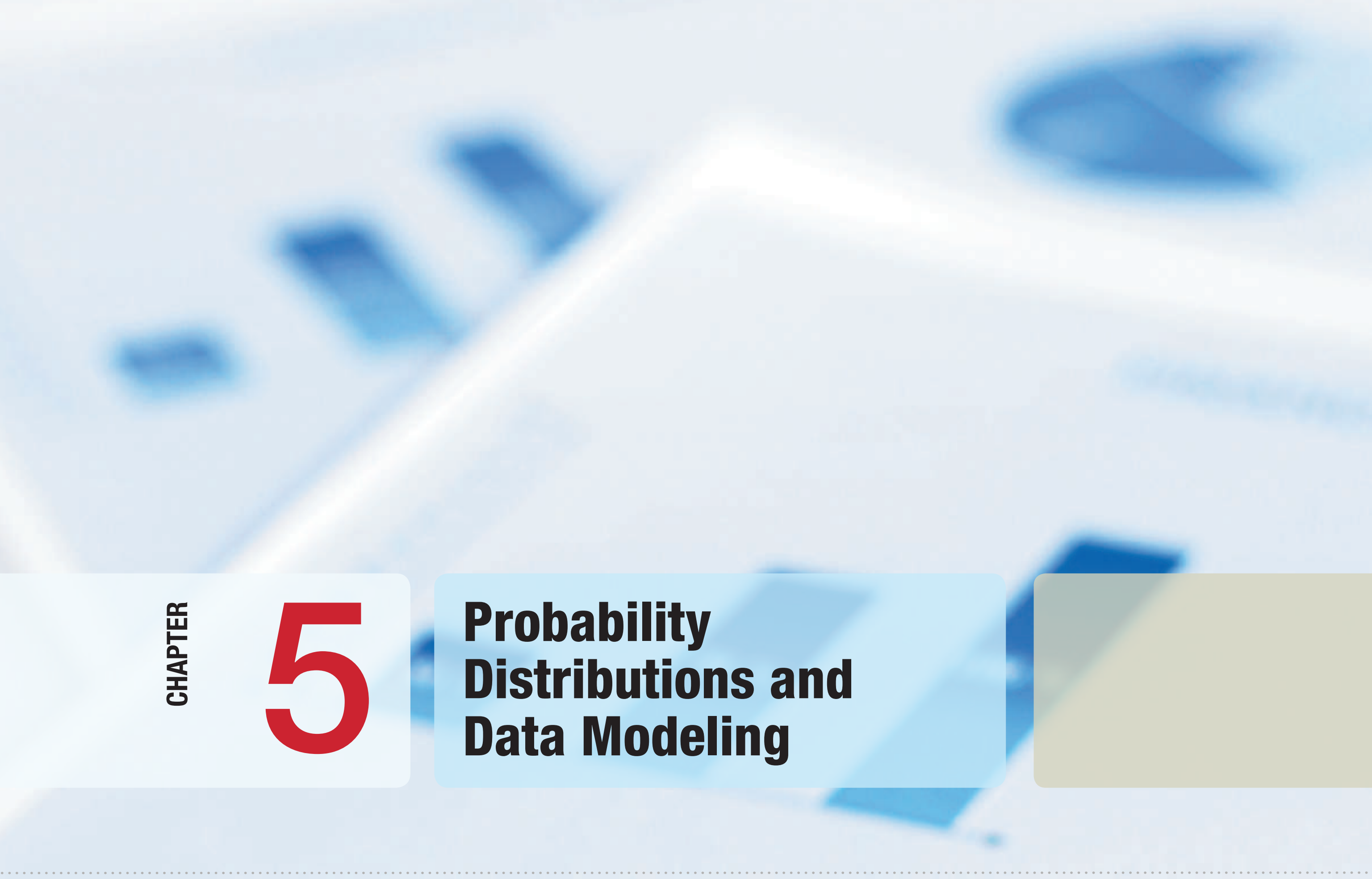
CHAPTER 5

Probability Distributions and Data Modeling



R-O-M-A/Shutterstock.com

Learning Objectives

After studying this chapter, you will be able to:

- Explain the concept of probability and provide examples of the three definitional perspectives of probability.
- Use probability rules and formulas to perform probability calculations.
- Explain conditional probability and how it can be applied in a business context.
- Compute conditional probabilities from cross-tabulation data.
- Determine if two events are independent using probability arguments.
- Apply the multiplication law of probability.
- Explain the difference between a discrete and a continuous random variable.
- Define a probability distribution.
- Verify the properties of a probability mass function.
- Use the cumulative distribution function to compute probabilities over intervals.
- Compute the expected value and variance of a discrete random variable.
- Use expected values to support simple business decisions.
- Calculate probabilities for the Bernoulli, binomial, and Poisson distributions, using the probability mass function and Excel functions.
- Explain how a probability density function differs from a probability mass function.
- List the key properties of probability density functions.
- Use the probability density and cumulative distribution functions to calculate probabilities for a uniform distribution.
- Describe the normal and standard normal distributions and use Excel functions to calculate probabilities.
- Use the standard normal distribution table and *z*-values to compute normal probabilities.

- Describe properties of the exponential distribution and compute probabilities.
- Give examples of other types of distributions used in business applications.
- Sample from discrete distributions in a spreadsheet using VLOOKUP.
- Use Excel's *Random Number Generation* tool.
- Generate random variates using *Analytic Solver Platform* functions.
- Fit distributions using *Analytic Solver Platform*.

Most business decisions involve some elements of uncertainty and randomness. For example, the times to repair computers in the *Computer Repair Times* Excel file that we discussed in Chapter 4 showed quite a bit of uncertainty that we needed to understand to provide information to customers about their computer repairs. We also saw that different samples of repair times result in different means, variances, and frequency distributions. Therefore, it would be beneficial to be able to identify some general characteristics of repair times that would apply to the entire population—even those repairs that have not yet taken place. In other situations, we may not have any data for analysis and simply need to make some judgmental assumptions about future uncertainties. For example, to develop a model to predict the profitability of a new and innovative product, we would need to make reliable assumptions about sales and consumer behavior without any prior data on which to base them. Characterizing the nature of distributions of data and specifying uncertain assumptions in decision models relies on fundamental knowledge of probability concepts and probability distributions—the subject of this chapter.

Basic Concepts of Probability

The notion of probability is used everywhere, both in business and in our daily lives; from market research and stock market predictions to the World Series of Poker and weather forecasts. In business, managers need to know such things as the likelihood that a new product will be profitable or the chances that a project will be completed on time. Probability quantifies the uncertainty that we encounter all around us and is an important building block for business analytics applications. **Probability** is the likelihood that an outcome—such as whether a new product will be profitable or not or whether a project will be completed within 15 weeks—occurs. Probabilities are expressed as values between 0 and 1, although many people convert them to percentages. The statement that there is a 10% chance that oil prices will rise next quarter is another way of stating that the probability of a rise in oil prices is 0.1. The closer the probability is to 1, the more likely it is that the outcome will occur.

To formally discuss probability, we need some new terminology. An **experiment** is a process that results in an outcome. An experiment might be as simple as rolling two dice, observing and recording weather conditions, conducting a market research study, or watching the stock market. The **outcome** of an experiment is a result that we observe; it might be the sum of two dice, a description of the weather, the proportion of consumers who favor a new product, or the change in the Dow Jones Industrial Average (DJIA) at the end of a week. The collection of all possible outcomes of an experiment is called the **sample space**. For instance, if we roll two fair dice, the possible outcomes are the numbers 2 through 12; if we observe the weather, the outcome might be clear, partly cloudy, or cloudy; the outcomes for customer reaction to a new product in a market research study would be favorable or unfavorable, and the weekly change in the DJIA can theoretically be any positive or negative real number. Note that a sample space may consist of a small number of discrete outcomes or an infinite number of outcomes.

Probability may be defined from one of three perspectives. First, if the process that generates the outcomes is known, probabilities can be deduced from theoretical arguments; this is the *classical definition* of probability.

EXAMPLE 5.1 Classical Definition of Probability

Suppose we roll two dice. If we examine all possible outcomes that may occur, we can easily determine that there are 36: rolling one of six numbers on the first die and rolling one of six numbers on the second die, for example, (1,1), (1,2), (1,3), . . . , (6,4), (6,5), (6,6). Out of these 36 possible outcomes, 1 outcome will be the number 2, 2 outcomes will be the number 3 (you can roll a 1 on the first die and 2 on the second, and vice versa), 6 outcomes will be the number 7, and so on. Thus, the probability of rolling any number is the ratio of the number of ways of rolling that number to the total number of possible outcomes. For instance, the probability of rolling a 2 is 1/36, the probability of rolling a 3 is 2/36 = 1/18, and the probability of rolling a 7 is 6/36 = 1/6. Similarly, if two consumers are asked whether or not they like a new product, there could be 4 possible outcomes:

1. (like, like)
2. (like, dislike)
3. (dislike, like)
4. (dislike, dislike)

If these are assumed to be equally likely, the probability that *at least* one consumer would respond unfavorably is 3/4.

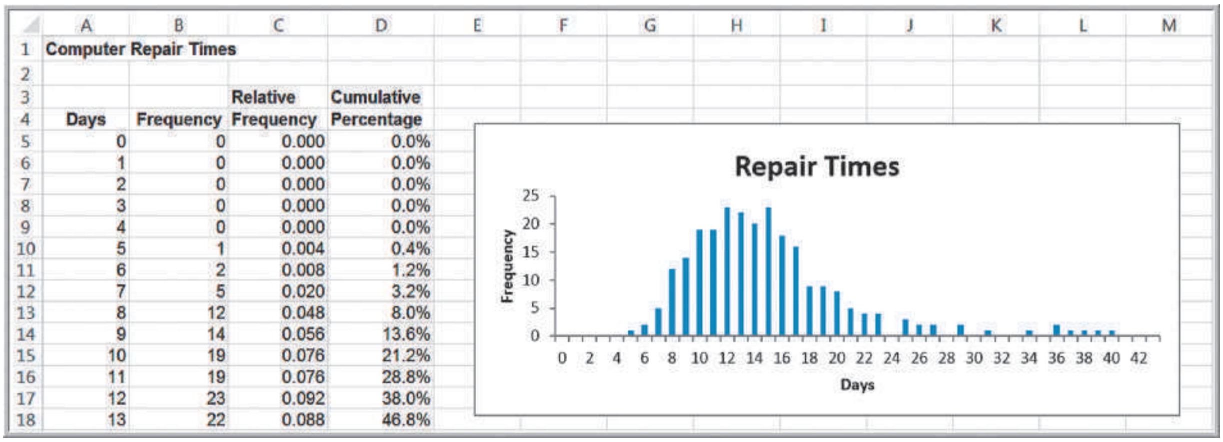
The second approach to probability, called the *relative frequency definition*, is based on empirical data. The probability that an outcome will occur is simply the relative frequency associated with that outcome.

EXAMPLE 5.2 Relative Frequency Definition of Probability

Using the sample of computer repair times in the Excel file *Computer Repair Times*, we developed the relative frequency distribution in Chapter 4, shown again in Figure 5.1. We could state that the probability that a computer would be repaired in as little as 4 days is 0, the probability that it would be repaired in exactly 10 days is 0.076, and so on. In using the relative frequency definition, it is important to understand that as more data become available, the distribution of outcomes and, hence, the probabilities may change.

Finally, the *subjective definition* of probability is based on judgment and experience, as financial analysts might use in predicting a 75% chance that the DJIA will increase 10% over the next year, or as sports experts might predict, at the start of the football season, a 1-in-5 chance (0.20 probability) of a certain team making it to the Super Bowl.

Which definition to use depends on the specific application and the information we have available. We will see various examples that draw upon each of these perspectives.



Computer Repair Times

Days	Frequency	Relative Frequency	Cumulative Percentage
0	0	0.000	0.0%
1	0	0.000	0.0%
2	0	0.000	0.0%
3	0	0.000	0.0%
4	0	0.000	0.0%
5	1	0.004	0.4%
6	2	0.008	1.2%
7	5	0.020	3.2%
8	12	0.048	8.0%
9	14	0.056	13.6%
10	19	0.076	21.2%
11	19	0.076	28.8%
12	23	0.092	38.0%
13	22	0.088	46.8%

Figure 5.1

Distribution of *Computer Repair Times*

Probability Rules and Formulas

Suppose we label the n outcomes in a sample space as $O_1, O_2, \ldots, O_n$, where O_i represents the ith outcome in the sample space. Let $P(O_i)$ be the probability associated with the outcome O_i. Two basic facts govern probability:

- The probability associated with any outcome must be between 0 and 1, or

$$0 \le P(O_i) \le 1 \text{ for each outcome } O_i \tag{5.1}$$

- The sum of the probabilities over all possible outcomes must be 1.0, or

$$P(O_1) + P(O_2) + \cdots + P(O_n) = 1 \tag{5.2}$$

An **event** is a collection of one or more outcomes from a sample space. An example of an event would be rolling a 7 or an 11 with two dice, completing a computer repair in between 7 and 14 days, or obtaining a positive weekly change in the DJIA. This leads to the following rule:

Rule 1. The probability of any event is the sum of the probabilities of the outcomes that comprise that event.

EXAMPLE 5.3 Computing the Probability of an Event

Consider the event of rolling a 7 or 11 on two dice. The probability of rolling a 7 is $\frac{6}{36}$ and the probability of rolling an 11 is $\frac{2}{36}$; thus, the probability of rolling a 7 or 11 is $\frac{6}{36} + \frac{2}{36} = \frac{8}{36}$. Similarly, the probability of repairing a computer in 7 days or less is the sum of the probabilities of the outcomes $O_1 = 0, O_2 = 1, O_3 = 2, O_4 = 3, O_5 = 4, O_6 = 5, O_7 = 6$, and $O_8 = 7$ days, or $P(O_6) + P(O_7) + P(O_8) = 0.004 + 0.008 + 0.020 = 0.032$ (note that the probabilities $P(O_1) = P(O_2) = P(O_3) = P(O_4) = P(O_5) = 0$; see Figure 5.1).

If A is any event, the **complement** of A, denoted A^c, consists of all outcomes in the sample space not in A.

Rule 2. The probability of the complement of any event A is $P(A^c) = 1 - P(A)$.

EXAMPLE 5.4 Computing the Probability of the Complement of an Event

If $A = \{7,11\}$ in the dice example, then $A^c = \{2, 3, 4, 5, 6, 8, 9, 10, 12\}$. Thus, the probability of rolling anything other than a 7 or 11 is $P(A^c) = 1 - \frac{8}{36} = \frac{28}{36}$. If $A = \{0, 1, 2, 3, 4, 5, 6, 7\}$ in the computer repair example, $A^c = \{8, 9, \ldots, 42\}$ and $P(A^c) = 1 - 0.032 = 0.968$. This is the probability of completing the repair in more than a week.

The **union** of two events contains all outcomes that belong to either of the two events. To illustrate this with rolling two dice, let A be the event $\{7, 11\}$ and B be the event $\{2, 3, 12\}$. The union of A and B is the event $\{2, 3, 7, 11, 12\}$. If A and B are two events, the probability that some outcome in either A or B (i.e., the union of A and B) occurs is denoted as $P(A \text{ or } B)$. Finding this probability depends on whether the events are mutually exclusive or not.

Two events are **mutually exclusive** if they have no outcomes in common. The events A and B in the dice example are mutually exclusive. When events are mutually exclusive, the following rule applies:

Rule 3. If events A and B are mutually exclusive, then $P(A \text{ or } B) = P\,(A) + P(B)$.

EXAMPLE 5.5 Computing the Probability of Mutually Exclusive Events

For the dice example, the probability of event $A = \{7, 11\}$ is $P(A) = \frac{8}{36}$, and the probability of event $B = \{2, 3, 12\}$ is $P(B) = \frac{4}{36}$. Therefore, the probability that either event A or B occurs, that is, the roll of the dice is either 2, 3, 7, 11, or 12, is $\frac{8}{36} + \frac{4}{36} = \frac{12}{36}$.

If two events are *not* mutually exclusive, then adding their probabilities would result in double-counting some outcomes, so an adjustment is necessary. This leads to the following rule:

Rule 4. If two events A and B are not mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.

Here, (A and B) represents the **intersection** of events A and B—that is, all outcomes belonging to both A and B.

EXAMPLE 5.6 Computing the Probability of Non–Mutually Exclusive Events

In the dice example, let us define the events $A = \{2, 3, 12\}$ and $B = \{\text{even number}\}$. Then A and B are not mutually exclusive because both events have the numbers 2 and 12 in common. Thus, the intersection (A and B) $= \{2, 12\}$. Therefore, $P(A \text{ or } B) = P\{2, 3, 12\} + P(\text{even number}) - P(A \text{ and } B) = \frac{4}{36} + \frac{18}{36} - \frac{2}{36} = \frac{20}{36}$.

Joint and Marginal Probability

In many applications, more than one event occurs simultaneously, or in statistical terminology, *jointly*. We will only discuss the simple case of two events. For instance, suppose that a sample of 100 individuals were asked to evaluate their preference for three new proposed energy drinks in a blind taste test. The sample space consists of two types of outcomes corresponding to each individual: gender (F = female or M = male) and brand preference (B_1, B_2, or B_3). We may define a new sample space consisting of the outcomes that reflect the different combinations of outcomes from these two sample spaces. Thus, for any respondent in the blind taste test, we have six possible (mutually exclusive) combinations of outcomes:

1. O_1 = the respondent is female and prefers brand 1
2. O_2 = the respondent is female and prefers brand 2
3. O_3 = the respondent is female and prefers brand 3
4. O_4 = the respondent is male and prefers brand 1
5. O_5 = the respondent is male and prefers brand 2
6. O_6 = the respondent is male and prefers brand 3

Here, the probability of each of these events is the intersection of the gender and brand preference event. For example, $P(O_1) = P(F \text{ and } B_1)$, $P(O_2) = P(F \text{ and } B_2)$, and so on. The probability of the intersection of two events is called a **joint probability**. The probability of an event, irrespective of the outcome of the other joint event, is called a **marginal probability**. Thus, $P(F)$, $P(M)$, $P(B_1)$, $P(B_2)$, and $P(B_3)$ would be marginal probabilities.

EXAMPLE 5.7 Applying Probability Rules to Joint Events

Figure 5.2 shows a portion of the data file *Energy Drink Survey*, along with a cross-tabulation constructed from a PivotTable. The joint probabilities of gender and brand preference are easily calculated by dividing the number of respondents corresponding to each of the six outcomes listed above by the total number of respondents, 100. Thus, $P(F \text{ and } B_1) = P(O_1) = 9/100 = 0.09$, $P(F \textit{ and } B_2) = P(O_2) = 6/100 = 0.06$, and so on. Note that the sum of the probabilities of all these outcomes is 1.0.

We see that the event F, (respondent is female) is comprised of the outcomes O_1, O_2, and O_3, and therefore $P(F) = P(O_1) + P(O_2) + P(O_3) = 0.37$ using Rule 1. The complement of this event is M; that is, the respondent is male. Note that $P(M) = 0.63 = 1 - P(F)$, as reflected by Rule 2. The event B_1 is comprised of the outcomes O_1 and O_4, and thus, $P(B_1) = P(O_1) + P(O_4) = 0.34$. Similarly, we find that $P(B2) = 0.23$ and $P(B_3) = 0.43$.

Events F and M are mutually exclusive, as are events B_1, B_2, and B_3 since a respondent may be only male or female and prefer exactly one of the three brands. We can use Rule 3 to find, for example, $P(B_1 \text{ or } B_2) = 0.34 + 0.23 = 0.57$. Events F and B_1, however, are not mutually exclusive because a respondent can be both female and prefer brand 1. Therefore, using Rule 4, we have $P(F \text{ or } B_1) = P(F) + P(B_1) - P(F \text{ and } B_1) = 0.37 + 0.34 - 0.09 = 0.62$.

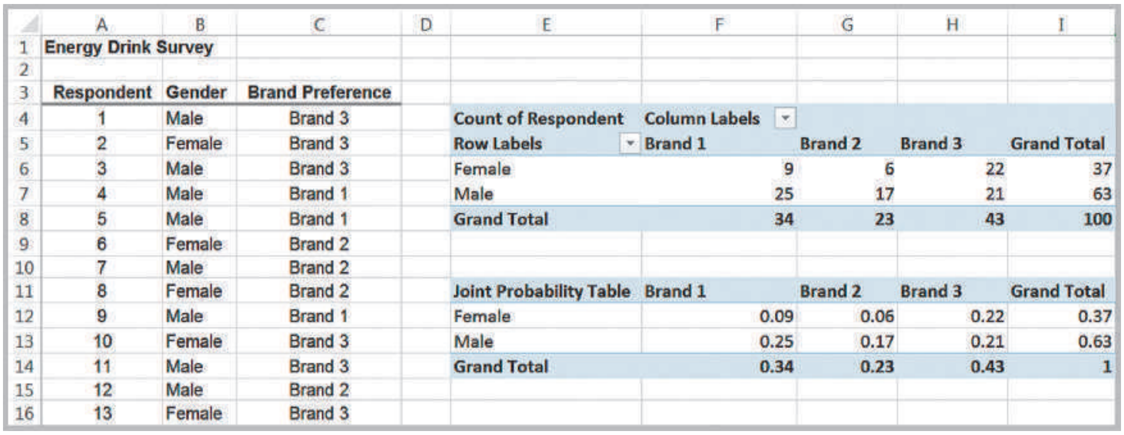
The joint probabilities can easily be computed, as we have seen, by dividing the values in the cross-tabulation by the total, 100. Below the PivotTable in Figure 5.2 is a **joint probability table**, which summarizes these joint probabilities.

The marginal probabilities are given in the margins of the joint probability table by summing the rows and columns. Note, for example, that $P(F) = P(F \text{ and } B_1) + P(F \text{ and } B_2) + P(F \text{ and } B_3) = 0.09 + 0.06 + 0.22 = 0.37$. Similarly, $P(B_1) = P(F \text{ and } B_1) + P(M \text{ and } B_1) = 0.09 + 0.25 = 0.34$.

This discussion of joint probabilities leads to the following probability rule:

Rule 5. If event A is comprised of the outcomes $\{A_1, A_2, \ldots, A_n\}$ and event B is comprised of the outcomes $\{B_1, B_2, \ldots, B_n\}$, then

$$P(A_i) = P(A_i \text{ and } B_1) + P(A_i \text{ and } B_2) + \cdots + P(A_i \text{ and } B_n)$$



	A	B	C	D	E	F	G	H	I
1	Energy Drink Survey								
2									
3	Respondent	Gender	Brand Preference						
4	1	Male	Brand 3		Count of Respondent	Column Labels ▾			
5	2	Female	Brand 3		Row Labels ▾	Brand 1	Brand 2	Brand 3	Grand Total
6	3	Male	Brand 3		Female	9	6	22	37
7	4	Male	Brand 1		Male	25	17	21	63
8	5	Male	Brand 1		Grand Total	34	23	43	100
9	6	Female	Brand 2						
10	7	Male	Brand 2						
11	8	Female	Brand 2		Joint Probability Table	Brand 1	Brand 2	Brand 3	Grand Total
12	9	Male	Brand 1		Female	0.09	0.06	0.22	0.37
13	10	Female	Brand 3		Male	0.25	0.17	0.21	0.63
14	11	Male	Brand 3		Grand Total	0.34	0.23	0.43	1
15	12	Male	Brand 2						
16	13	Female	Brand 3						

Figure 5.2

Portion of Excel File *Energy Drink Survey*

Conditional Probability

Conditional probability is the probability of occurrence of one event A, given that another event B is known to be true or has already occurred.

EXAMPLE 5.8 Computing a Conditional Probability in a Cross-Tabulation

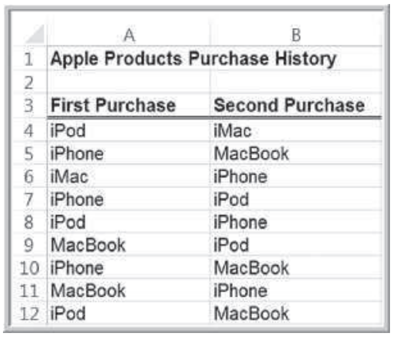
We will use the information shown in the energy drink survey example in Figure 5.2 to illustrate how to compute conditional probabilities from a cross-tabulation or joint probability table.

Suppose that we know that a respondent is male. What is the probability that he prefers brand 1? From the PivotTable, note that there are only 63 males in the group and of these, 25 prefer brand 1. Therefore, the probability that a male respondent prefers brand 1 is $\frac{25}{63}$. We could have obtained the same result from the joint probability table by dividing the joint probability 0.25 (the probability that the respondent is male and prefers brand 1) by the marginal probability 0.63 (the probability that the respondent is male).

Conditional probabilities are useful in analyzing data in cross-tabulations, as well as in other types of applications. Many companies save purchase histories of customers to predict future sales. Conditional probabilities can help to predict future purchases based on past purchases.

EXAMPLE 5.9 Conditional Probability in Marketing

The Excel file *Apple Purchase History* presents a hypothetical history of consumer purchases of Apple products, showing the first and second purchase for a sample of 200 customers that have made repeat purchases (see Figure 5.3). The PivotTable in Figure 5.4 shows the count of the type of second purchase given that each product was purchased first. For example, 13 customers purchased iMacs as their first Apple product. Then the conditional probability of purchasing an iPad given that the customer first purchased an iMac is $\frac{2}{13} = 0.15$. Similarly, 74 customers purchased a MacBook as their first purchase; the conditional probability of purchasing an iPhone if a customer first purchased a MacBook is $\frac{26}{74} = 0.35$. By understanding which products are more likely to be purchased by customers who already own other products, companies can better target advertising strategies.

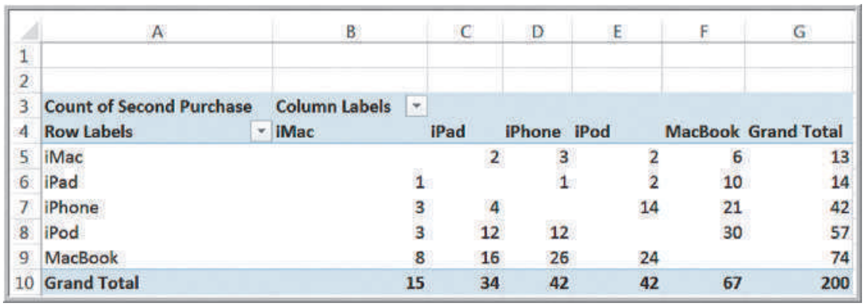


Apple Products Purchase History

First Purchase	Second Purchase
iPod	iMac
iPhone	MacBook
iMac	iPhone
iPhone	iPod
iPod	iPhone
MacBook	iPod
iPhone	MacBook
MacBook	iPhone
iPod	MacBook

Figure 5.3

Portion of Excel File *Apple Purchase History*



Count of Second Purchase	Column Labels					
Row Labels	iMac	iPad	iPhone	iPod	MacBook	Grand Total
iMac		2	3	2	6	13
iPad	1		1	2	10	14
iPhone	3	4		14	21	42
iPod	3	12	12		30	57
MacBook	8	16	26	24		74
Grand Total	15	34	42	42	67	200

Figure 5.4

PivotTable of Purchase Behavior

In general, the conditional probability of an event A given that event B is known to have occurred is

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{5.3}$$

We read the notation $P(A|B)$ as "the probability of A given B."

EXAMPLE 5.10 Using the Conditional Probability Formula

Using the data from the energy drink survey example, substitute B_1 for A and M for B in formula (5.3). This results in the conditional probability of B_1 given M:

$$P(B_1|M) = \frac{P(B_1 \text{ and } M)}{P(M)} = \frac{0.25}{0.63} = 0.397.$$

Similarly, the probability of preferring brand 1 if the respondent is female is

$$P(B_1|F) = \frac{P(B_1 \text{ and } F)}{P(F)} = \frac{0.09}{0.37} = 0.243.$$

The following table summarizes the conditional probabilities of brand preference given gender:

P(Brand|Gender)	Brand 1	Brand 2	Brand 3
Male	0.397	0.270	0.333
Female	0.243	0.162	0.595

Such information can be important in marketing efforts. Knowing that there is a difference in preference by gender can help focus advertising. For example, we see that about 40% of males prefer brand 1, whereas only about 24% of females do, and a higher proportion of females prefer brand 3. This suggests that it would make more sense to focus on advertising brand 1 more in male-oriented media and brand 3 in female-oriented media.

The conditional probability formula may be used in other ways. For example, multiplying both sides of formula (5.3) by $P(B)$, we obtain $P(A \text{ and } B) = P(A|B)\,P(B)$. Note that we may switch the roles of A and B and write $P(B \text{ and } A) = P(B|A)\,P(A)$. But $P(B \text{ and } A)$ is the same as $P(A \text{ and } B)$; thus we can express $P(A \text{ and } B)$ in two ways:

$$P(A \text{ and } B) = P(A|B)\,P(B) = P(B|A)\,P(A) \tag{5.4}$$

This is often called the **multiplication law of probability**.

We may use this concept to express the probability of an event in a joint probability table in a different way. Using the energy drink survey again in Figure 5.2, note that

$$P(F) = P(F \text{ and } \textit{Brand 1}) + P(F \text{ and } \textit{Brand 2}) + P(F \text{ and } \textit{Brand 3})$$

Using formula (5.4), we can express the joint probabilities $P(A \text{ and } B)$ by $P(A|B)\,P(B)$. Therefore,

$P(F) = P(F|\textit{Brand 1})\,P(\textit{Brand 1}) + P(F|\textit{Brand 2})\,P(\textit{Brand 2}) + P(F|\textit{Brand 3})\,P(\textit{Brand 3}) = (0.265)(0.34) + (0.261)(0.23) + (0.512)(0.43) = 0.37$ (within rounding precision).

We can express this calculation using the following extension of the multiplication law of probability. Suppose $B_1, B_2, \ldots, B_n$ are mutually exclusive events whose union comprises the entire sample space. Then

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n) \quad \textbf{(5.5)}$$

EXAMPLE 5.11 Using the Multiplication Law of Probability

Texas Hold 'Em has become a popular game because of the publicity surrounding the World Series of Poker. At the beginning of a game, players each receive two cards face down (we won't worry about how the rest of the game is played). Suppose that a player receives an ace on her first card. The probability that she will end up with "pocket aces" (two aces in the hand) is *P*(*ace on first card and ace on second card*) = *P*(*ace on second card* | *ace on first card*) × *P*(*ace on first card*). Since the probability of an ace on the first card is 4/52 and the probability of an ace on the second card if she has already drawn an ace is 3/51, we have

$$\begin{aligned} &P(\text{ace on first card and ace on second card}) \\ &\qquad = P(\text{ace on second card}\,|\,\text{ace on first card}) \\ &\qquad\quad \times \text{P(ace on first card)} \\ &\qquad = \left(\frac{3}{51}\right) \times \left(\frac{4}{52}\right) = 0.004525 \end{aligned}$$

In Example 5.10, we see that the probability of preferring a brand depends on gender. We may say that brand preference and gender are not independent. We may formalize this concept by defining the notion of **independent events**: *Two events A and B are independent if* $P(A|B) = P(A)$.

EXAMPLE 5.12 Determining if Two Events Are Independent

We use this definition in the energy drink survey example. Recall that the conditional probabilities of brand preference given gender are

P(Brand | Gender)	Brand 1	Brand 2	Brand 3
Male	0.397	0.270	0.333
Female	0.243	0.162	0.595

We see that whereas $P(B_1|M) = 0.397$, $P(B_1)$ was shown to be 0.34 in Example 5.7; thus, these two events are not independent.

Finally, we see that if two events are independent, then we can simplify the multiplication law of probability in equation (5.4) by substituting $P(A)$ for $P(A|B)$:

$$P(A \text{ and } B) = P(B)\,P(A) = P(A)P(B) \quad \textbf{(5.6)}$$

EXAMPLE 5.13 Using the Multiplication Law for Independent Events

Suppose *A* is the event that a 6 is first rolled on a pair of dice and *B* is the event of rolling a 2, 3, or 12 on the next roll. These events are independent because the roll of a pair of dice does not depend on the previous roll. Then we may compute $P(A \text{ and } B) = P(A)P(B) = \left(\frac{5}{36}\right)\left(\frac{4}{36}\right) = \frac{20}{1296}$.

Random Variables and Probability Distributions

Some experiments naturally have numerical outcomes, such as a roll of the dice, the time it takes to repair computers, or the weekly change in a stock market index. For other experiments, such as obtaining consumer response to a new product, the sample space is categorical. To have a consistent mathematical basis for dealing with probability, we would like the outcomes of all experiments to be numerical. A **random variable** is a numerical description of the outcome of an experiment. Formally, a random variable is a function that assigns a real number to each element of a sample space. If we have categorical outcomes, we can associate an arbitrary numerical value to them. For example, if a consumer likes a product in a market research study, we might assign this outcome a value of 1; if the consumer dislikes the product, we might assign this outcome a value of 0. Random variables are usually denoted by capital italic letters, such as *X* or *Y*.

Random variables may be discrete or continuous. A **discrete random variable** is one for which the number of possible outcomes can be counted. A **continuous random variable** has outcomes over one or more continuous intervals of real numbers.

EXAMPLE 5.14 Discrete and Continuous Random Variables

The outcomes of rolling two dice (the numbers 2 through 12) and customer reactions to a product (like or dislike) are discrete random variables. The number of outcomes may be finite or theoretically infinite, such as the number of hits on a Web site link during some period of time—we cannot place a guaranteed upper limit on this number; nevertheless, the number of hits can be counted. Example of continuous random variables are the weekly change in the DJIA, which may assume any positive or negative value, the daily temperature, the time to complete a task, the time between failures of a machine, and the return on an investment.

A **probability distribution** is the characterization of the possible values that a random variable may assume along with the probability of assuming these values. A probability distribution can be either discrete or continuous, depending on the nature of the random variable it models. Discrete distributions are easier to understand and work with, and we deal with them first.

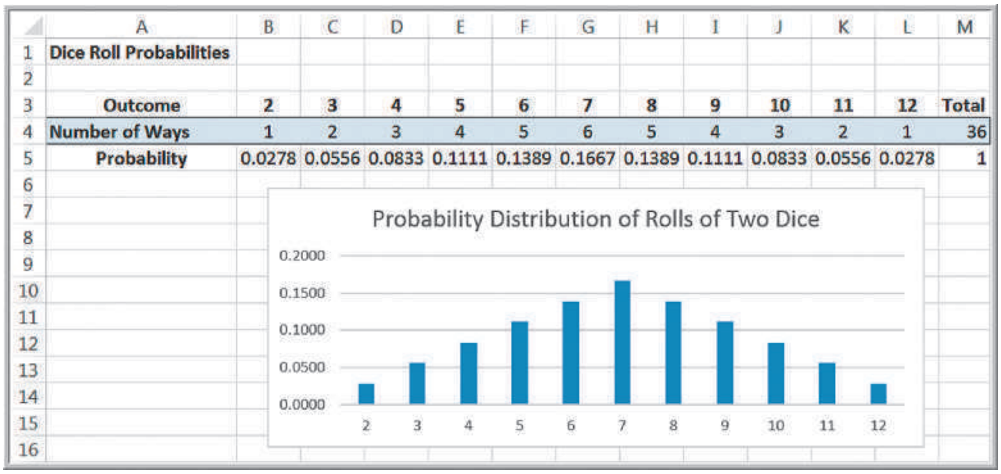
We may develop a probability distribution using any one of the three perspectives of probability. First, if we can quantify the probabilities associated with the values of a random variable from theoretical arguments; then we can easily define the probability distribution.

EXAMPLE 5.15 Probability Distribution of Dice Rolls

The probabilities of the outcomes for rolling two dice are calculated by counting the number of ways to roll each number divided by the total number of possible outcomes. These, along with an Excel column chart depicting the probability distribution, are shown from the Excel file *Dice Rolls* in Figure 5.5.

Figure 5.5

Probability Distribution of Rolls of Two Dice



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Dice Roll Probabilities												
2													
3	Outcome	2	3	4	5	6	7	8	9	10	11	12	Total
4	Number of Ways	1	2	3	4	5	6	5	4	3	2	1	36
5	Probability	0.0278	0.0556	0.0833	0.1111	0.1389	0.1667	0.1389	0.1111	0.0833	0.0556	0.0278	1

Second, we can calculate the relative frequencies from a sample of empirical data to develop a probability distribution. Thus, the relative frequency distribution of computer repair times (Figure 5.1) is an example. Because this is based on sample data, we usually call this an **empirical probability distribution**. An empirical probability distribution is an approximation of the probability distribution of the associated random variable, whereas the probability distribution of a random variable, such as the one derived from counting arguments, is a theoretical model of the random variable.

Finally, we could simply specify a probability distribution using subjective values and expert judgment. This is often done in creating decision models for the phenomena for which we have no historical data.

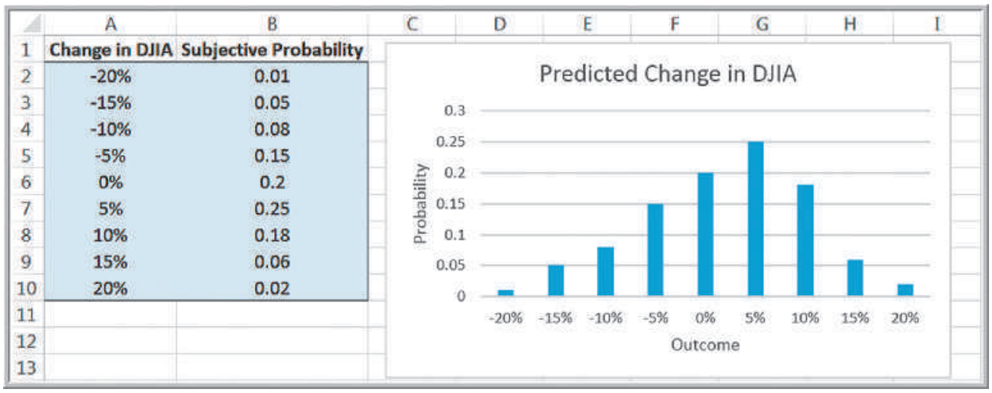
EXAMPLE 5.16 A Subjective Probability Distribution

Figure 5.6 shows a hypothetical example of the distribution of one expert's assessment of how the DJIA might change in the next year. This might have been created purely by intuition and expert judgment, but we hope it would be supported by some extensive analysis of past and current data using business analytics tools.

Researchers have identified many common types of probability distributions that are useful in a variety of applications of business analytics. A working knowledge of common families of probability distributions is important for several reasons. First, it can help you to understand the underlying process that generates sample data. We investigate the relationship between distributions and samples later. Second, many phenomena in business and nature follow some theoretical distribution and, therefore, are useful in building decision models. Finally, working with distributions is essential in computing probabilities of occurrence of outcomes to assess risk and make decisions.

Figure 5.6

Subjective Probability Distribution of DJIA Change



	A	B
1	Change in DJIA	Subjective Probability
2	-20%	0.01
3	-15%	0.05
4	-10%	0.08
5	-5%	0.15
6	0%	0.2
7	5%	0.25
8	10%	0.18
9	15%	0.06
10	20%	0.02
11		
12		
13		

Discrete Probability Distributions

For a discrete random variable X, the probability distribution of the discrete outcomes is called a **probability mass function** and is denoted by a mathematical function, $f(x)$. The symbol x_i represents the ith value of the random variable X and $f(x_i)$ is the probability.

EXAMPLE 5.17 Probability Mass Function for Rolling Two Dice

For instance, in Figure 5.5 for the dice example, the values of the random variable X, which represents the sum of the rolls of two dice, are $x_1 = 2$, $x_2 = 3$ $x_3 = 4$, $x_4 = 5$, $x_5 = 6$, $x_6 = 7$, $x_7 = 8$, $x_8 = 9$, $x_3 = 10$ $x_{10} = 11$, $x_{11} = 12$. The probability mass function for X is

$$f(x_1) = \frac{1}{36} = 0.0278$$

$$f(x_2) = \frac{2}{36} = 0.0556$$

$$f(x_3) = \frac{3}{36} = 0.0833$$

$$f(x_4) = \frac{4}{36} = 0.1111$$

$$f(x_5) = \frac{5}{36} = 0.1389$$

$$f(x_6) = \frac{6}{36} = 0.1667$$

$$f(x_7) = \frac{5}{36} = 0.1389$$

$$f(x_8) = \frac{4}{36} = 0.1111$$

$$f(x_9) = \frac{3}{36} = 0.0833$$

$$f(x_{10}) = \frac{2}{36} = 0.0556$$

$$f(x_{11}) = \frac{1}{36} = 0.0278$$

A probability mass function has the properties that (1) the probability of each outcome must be between 0 and 1 and (2) the sum of all probabilities must add to 1; that is,

$$0 \leq f(x_i) \leq 1 \quad \text{for all } i \tag{5.7}$$

$$\sum_i f(x_i) = 1 \tag{5.8}$$

You can easily verify that this holds in each of the examples we have described.

A **cumulative distribution function**, $F(x)$, specifies the probability that the random variable X assumes a value *less than or equal to* a specified value, x. This is also denoted as $P(X \leq x)$ and read as "the probability that the random variable X is less than or equal to x."

EXAMPLE 5.18 Using the Cumulative Distribution Function

The cumulative distribution function for rolling two dice is shown in Figure 5.7, along with an Excel line chart that describes it visually from the worksheet *CumDist* in the *Dice Rolls* Excel file. To use this, suppose we want to know the probability of rolling a 6 or less. We simply look up the cumulative probability for 6, which is 0.5833. Alternatively, we could locate the point for $x = 6$ in the chart and estimate the probability from the graph. Also note that since the probability of rolling a 6 or less is 0.5833, then the probability of the complementary event (rolling a 7 or more) is $1 - 0.5833 = 0.4167$. We can also use the cumulative distribution function to find probabilities over intervals. For example, to find the probability of rolling a number between 4 and 8, $P(4 \leq X \leq 8)$, we can find $P(X \leq 8)$ and subtract $P(X \leq 3)$; that is,

$$P(4 \leq X \leq 8) = P(X \leq 8) - P(X \leq 3)$$
$$= 0.7222 - 0.0833 = 0.6389.$$

A word of caution. Be careful with the endpoints when computing probabilities over intervals for discrete distributions; because 4 is included in the interval we wish to compute, we need to subtract $P(X \leq 3)$, not $P(X \leq 4)$.

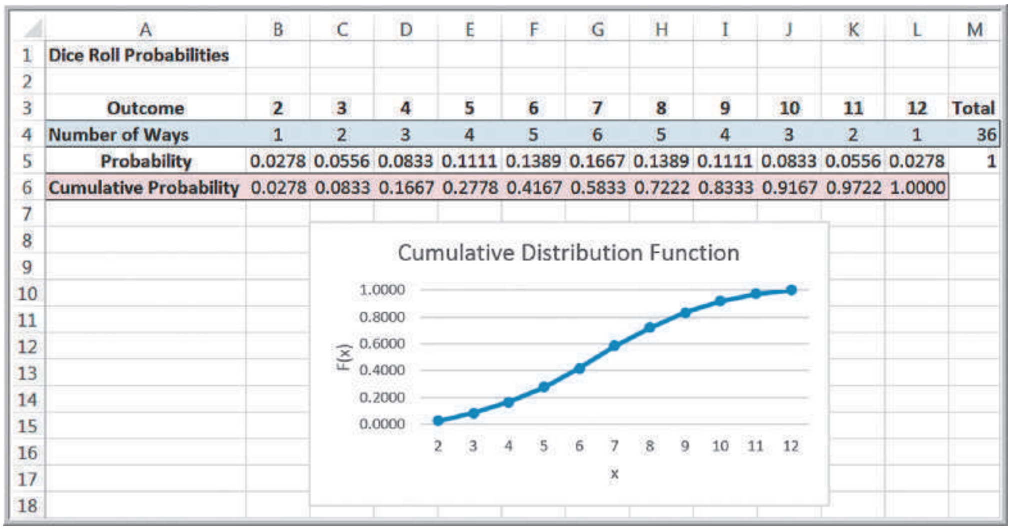
Expected Value of a Discrete Random Variable

The **expected value** of a random variable corresponds to the notion of the mean, or average, for a sample. For a discrete random variable X, the expected value, denoted $E[X]$, is the weighted average of all possible outcomes, where the weights are the probabilities:

$$E[X] = \sum_{i=1}^{\infty} x_i f(x_i) \qquad \textbf{(5.9)}$$

Figure 5.7

Cumulative Distribution Function for Rolling Two Dice



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Dice Roll Probabilities												
2													
3	Outcome	2	3	4	5	6	7	8	9	10	11	12	Total
4	Number of Ways	1	2	3	4	5	6	5	4	3	2	1	36
5	Probability	0.0278	0.0556	0.0833	0.1111	0.1389	0.1667	0.1389	0.1111	0.0833	0.0556	0.0278	1
6	Cumulative Probability	0.0278	0.0833	0.1667	0.2778	0.4167	0.5833	0.7222	0.8333	0.9167	0.9722	1.0000	

Note the similarity to computing the population mean using formula (4.13) in Chapter 4:

$$\mu = \frac{\sum_{i=1}^{N} f_i x_i}{N}$$

If we write this as the sum of x_i multiplied by (f_i/N), then we can think of f_i/N as the probability of x_i. Then this expression for the mean has the same basic form as the expected value formula.

EXAMPLE 5.19 Computing the Expected Value

We may apply formula (5.9) to the probability distribution of rolling two dice. We multiply the outcome 2 by its probability 1/36, add this to the product of the outcome 3 and its probability, and so on. Continuing in this fashion, the expected value is

$$\begin{aligned} E[X] = {} & 2(0.0278) + 3(0.0556) + 4(0.0833) + 5(0.01111) \\ & + 6(0.1389) + 7(0.1667) + 8(0.1389) + 9(0.1111) \\ & + 10(0.0833) + 11(0.0556) + 12(0.0278) = 7 \end{aligned}$$

Figure 5.8 shows these calculations in an Excel spreadsheet (worksheet *Expected Value* in the *Dice Rolls* Excel file). As expected (no pun intended), the average value of the roll of two dice is 7.

Using Expected Value in Making Decisions

Expected value can be helpful in making a variety of decisions, even those we see in daily life.

EXAMPLE 5.20 Expected Value on Television

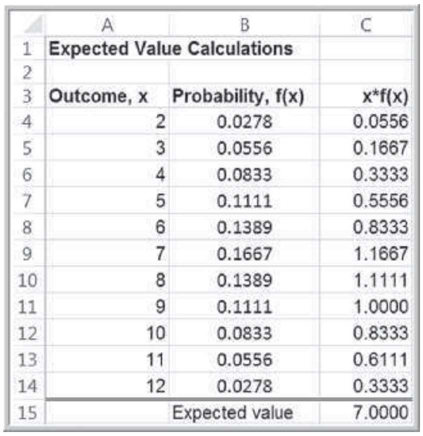
One of the author's favorite examples stemmed from a task in season 1 of Donald Trump's TV show, *The Apprentice*. Teams were required to select an artist and sell his or her art for the highest total amount of money. One team selected a mainstream artist who specialized in abstract art that sold for between $1,000 and $2,000; the second team chose an avant-garde artist whose surrealist and rather controversial art was priced much higher. Guess who won? The first team did, because the probability of selling a piece of mainstream art was much higher than the avant-garde artist whose bizarre art (the team members themselves didn't even like it!) had a very low probability of a sale. A back-of-the-envelope expected value calculation would have easily predicted the winner.

A popular game show that took TV audiences by storm several years ago was called *Deal or No Deal*. The game involved a set of numbered briefcases that contain amounts of money from 1 cent to $1,000,000. Contestants begin choosing cases to be opened and removed, and their amounts are shown. After each set of cases is opened, the banker offers the contestant an amount of money to quit the game, which the contestant may either choose or reject. Early in the game, the banker's offer is usually less than the expected value of the remaining cases, providing an incentive to continue. However, as the number of remaining cases becomes small, the banker's offers approach or may even exceed the average of the remaining cases. Most people press on until the bitter end and often walk away with a smaller amount than they could have had they been able to estimate the expected value of the remaining cases and make a more rational decision. In one case, a contestant had five briefcases left with $100, $400, $1,000, $50,000, and $300,000. Because the choice of each case is equally likely, the expected value was 0.2($100 + $400 + $1000 + $50,000 + $300,000) = $70,300 and the banker offered $80,000 to quit. Instead, she said "No Deal" and proceeded to open the $300,000 suitcase, eliminating it from the game, and took the next banker's offer of $21,000, which was more than 60% larger than the expected value of the remaining cases.[1]

[1] "Deal or No Deal: A Statistical Deal." www.pearsonified.com/2006/03/deal_or_no_deal_the_real_deal.php

Figure 5.8

Expected Value Calculations for Rolling Two Dice



	A	B	C
1	Expected Value Calculations		
2			
3	Outcome, x	Probability, f(x)	x*f(x)
4	2	0.0278	0.0556
5	3	0.0556	0.1667
6	4	0.0833	0.3333
7	5	0.1111	0.5556
8	6	0.1389	0.8333
9	7	0.1667	1.1667
10	8	0.1389	1.1111
11	9	0.1111	1.0000
12	10	0.0833	0.8333
13	11	0.0556	0.6111
14	12	0.0278	0.3333
15		Expected value	7.0000

It is important to understand that the expected value is a "long-run average" and is appropriate for decisions that occur on a repeated basis. For one-time decisions, however, you need to consider the downside risk and the upside potential of the decision. The following example illustrates this.

EXAMPLE 5.21 Expected Value of a Charitable Raffle

Suppose that you are offered the chance to buy one of 1,000 tickets sold in a charity raffle for $50, with the prize being $25,000. Clearly, the probability of winning is $\frac{1}{1,000}$, or 0.001, whereas the probability of losing is 1 – 0.001 – 0.999. The random variable X is your net winnings, and its probability distribution is

x	$f(x)$
−$50	0.999
$24,950	0.001

The expected value, $E[X]$, *is* −$50(0.999) + $24,950(0.001) = −$25.00. This means that if you played this game repeatedly over the long run, you would lose an average of $25.00 *each time* you play. Of course, for any *one* game, you would either lose $50 or win $24,950. So the question becomes, Is the risk of losing $50 worth the potential of winning $24,950? Although the expected value is negative, you might take the chance because the upside potential is large relative to what you might lose, and, after all, it is for charity. However, if your potential loss is large, you might not take the chance, even if the expected value were positive.

Decisions based on expected values are common in real estate development, day trading, and pharmaceutical research projects. Drug development is a good example. The cost of research and development projects in the pharmaceutical industry is generally in the hundreds of millions of dollars and often approaches $1 billion. Many projects never make it to clinical trials or might not get approved by the Food and Drug Administration. Statistics indicate that 7 of 10 products fail to return the cost of the company's capital. However, large firms can absorb such losses because the return from one or two blockbuster drugs can easily offset these losses. On an average basis, drug companies make a net profit from these decisions.

EXAMPLE 5.22 Airline Revenue Management

Let us consider a simplified version of the typical revenue management process that airlines use. At any date prior to a scheduled flight, airlines must make a decision as to whether to reduce ticket prices to stimulate demand for unfilled seats. If the airline does not discount the fare, empty seats might not be sold and the airline will lose revenue. If the airline discounts the remaining seats too early (and could have sold them at the higher fare), they would lose profit. The decision depends on the probability p of selling a full-fare ticket if they choose not to discount the price. Because an airline makes hundreds or thousands of such decisions each day, the expected value approach is appropriate.

Assume that only two fares are available: full and discount. Suppose that a full-fare ticket is \$560, the discount fare is \$400, and $p = 0.75$. For simplification, assume that if the price is reduced, then any remaining seats would be sold at that price. The expected value of not discounting the price is 0.25 (0) + 0.75(\$560) = \$420. Because this is higher than the discounted price, the airline should not discount at this time. In reality, airlines constantly update the probability p based on the information they collect and analyze in a database. When the value of p drops below the break-even point: \$400 = p(\$560), *or* $p = 0.714$, then it is beneficial to discount. It can also work in reverse; if demand is such that the probability that a higher-fare ticket would be sold, then the price may be adjusted upward. This is why published fares constantly change and why you may receive last-minute discount offers or may pay higher prices if you wait too long to book a reservation. Other industries such as hotels and cruise lines use similar decision strategies.

Variance of a Discrete Random Variable

We may compute the variance, Var[X], of a discrete random variable X as a weighted average of the squared deviations from the expected value:

$$\text{Var}[X] = \sum_{j=1}^{\infty} (x_j - E[X])^2 f(x_j) \qquad \textbf{(5.10)}$$

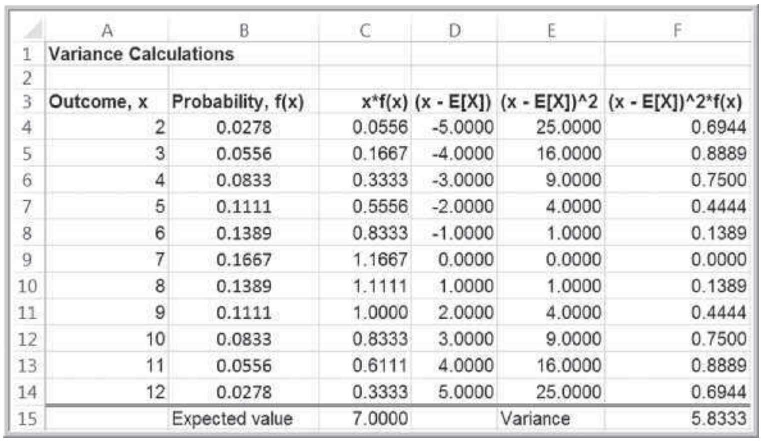
EXAMPLE 5.23 Computing the Variance of a Random Variable

We may apply formula (5.10) to calculate the variance of the probability distribution of rolling two dice. Figure 5.9 shows these calculations in an Excel spreadsheet (worksheet *Variance* in *Random Variable Calculations* Excel file).

Similar to our discussion in Chapter 4, the variance measures the uncertainty of the random variable; the higher the variance, the higher the uncertainty of the outcome. Although variances are easier to work with mathematically, we usually measure the variability of a random variable by its standard deviation, which is simply the square root of the variance.

Figure 5.9

Variance Calculations for Rolling Two Dice



	A	B	C	D	E	F
1	Variance Calculations					
2						
3	Outcome, x	Probability, f(x)	x*f(x)	(x - E[X])	(x - E[X])^2	(x - E[X])^2*f(x)
4	2	0.0278	0.0556	-5.0000	25.0000	0.6944
5	3	0.0556	0.1667	-4.0000	16.0000	0.8889
6	4	0.0833	0.3333	-3.0000	9.0000	0.7500
7	5	0.1111	0.5556	-2.0000	4.0000	0.4444
8	6	0.1389	0.8333	-1.0000	1.0000	0.1389
9	7	0.1667	1.1667	0.0000	0.0000	0.0000
10	8	0.1389	1.1111	1.0000	1.0000	0.1389
11	9	0.1111	1.0000	2.0000	4.0000	0.4444
12	10	0.0833	0.8333	3.0000	9.0000	0.7500
13	11	0.0556	0.6111	4.0000	16.0000	0.8889
14	12	0.0278	0.3333	5.0000	25.0000	0.6944
15		Expected value	7.0000		Variance	5.8333

Bernoulli Distribution

The **Bernoulli distribution** characterizes a random variable having two possible outcomes, each with a constant probability of occurrence. Typically, these outcomes represent "success" ($x = 1$) having probability p and "failure" ($x = 0$), having probability $1 - p$. A success can be any outcome you define. For example, in attempting to boot a new computer just off the assembly line, we might define a success as "does not boot up" in defining a Bernoulli random variable to characterize the probability distribution of a defective product. Thus, success need not be a favorable result in the traditional sense.

The probability mass function of the Bernoulli distribution is

$$f(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases} \tag{5.11}$$

where p represents the probability of success. The expected value is p, and the variance is $p(1 - p)$.

EXAMPLE 5.24 Using the Bernoulli Distribution

A Bernoulli distribution might be used to model whether an individual responds positively ($x = 1$) or negatively ($x = 0$) to a telemarketing promotion. For example, if you estimate that 20% of customers contacted will make a purchase, the probability distribution that describes whether or not a particular individual makes a purchase is Bernoulli with $p = 0.2$. Think of the following experiment. Suppose that you have a box with 100 marbles, 20 red and 80 white. For each customer, select one marble at random (and then replace it). The outcome will have a Bernoulli distribution. If a red marble is chosen, then that customer makes a purchase; if it is white, the customer does not make a purchase.

Binomial Distribution

The **binomial distribution** models n independent replications of a Bernoulli experiment, each with a probability p of success. The random variable X represents the number of successes in these n experiments. In the telemarketing example, suppose that we call $n = 10$ customers, each of which has a probability $p = 0.2$ of making a purchase. Then the probability distribution of the number of positive responses obtained from 10 customers is binomial. Using the binomial distribution, we can calculate the probability that exactly x customers out of the 10 will make a purchase for any value of x between 0 and 10. A binomial distribution might also be used to model the results of sampling inspection in a production operation or the effects of drug research on a sample of patients.

The probability mass function for the binomial distribution is

$$f(x) = \begin{cases} \binom{n}{x} p^{x}(1 - p)^{n-x}, & \text{for } x = 0, 1, 2, \ldots, n \\ 0, & \text{otherwise} \end{cases} \tag{5.12}$$

The notation $\binom{n}{x}$ represents the number of ways of choosing x distinct items from a group of n items and is computed as

$$\binom{n}{x} = \frac{n!}{x!\,(n - x)!} \tag{5.13}$$

where $n!$ (n factorial) $= n(n-1)(n-2)\cdots(2)(1)$, and $0!$ is defined to be 1.

EXAMPLE 5.25 Computing Binomial Probabilities

We may use formula (5.12) to compute binomial probabilities. For example, if the probability that any individual will make a purchase from a telemarketing solicitation is 0.2, then the probability distribution that *x* individuals out of 10 calls will make a purchase is

$$f(x) = \begin{cases} \binom{10}{x}(0.2)^x(0.8)^{10-x}, & \text{for } x = 0, 1, 2, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

Thus, to find the probability that 3 people will make a purchase among the 10 calls, we compute

$$f(3) = \binom{10}{3}(0.2)^3(0.8)^{10-3}$$

$$= (10!/3!7!)(0.008)(0.2097152)$$
$$= 120(0.008)(0.2097152) = 0.20133$$

The formula for the probability mass function for the binomial distribution is rather complex, and binomial probabilities are tedious to compute by hand; however, they can easily be computed in Excel using the function

BINOM.DIST(*number_s*, *trials*, *probability_s*, *cumulative*)

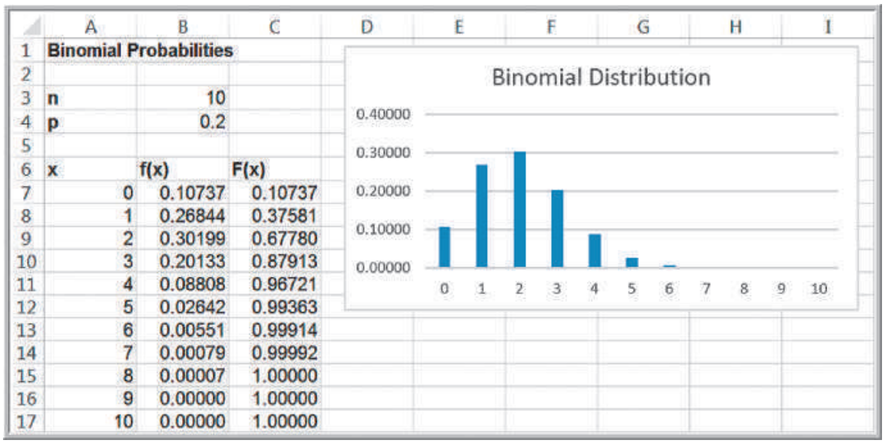
In this function, *number_s* plays the role of *x*, and *probability_s* is the same as *p*. If *cumulative* is set to TRUE, then this function will provide cumulative probabilities; otherwise the default is FALSE, and it provides values of the probability mass function, $f(x)$.

EXAMPLE 5.26 Using Excel's Binomial Distribution Function

Figure 5.10 shows the results of using this function to compute the distribution for the previous example (Excel file *Binomial Probabilities*). For instance, the probability that exactly 3 individuals will make a purchase is BINOM.DIST(A10,$B$3,$B$4,FALSE) = 0.20133 = *f*(3). The probability that 3 or fewer individuals will make a purchase is BINOM.DIST(A10,$B$3,$B$4,TRUE) = 0.87913 = *F*(3). Correspondingly, the probability that more than 3 out of 10 individuals will make a purchase is 1 − *F*(3) = 1 − 0.87913 = 0.12087.

Figure 5.10

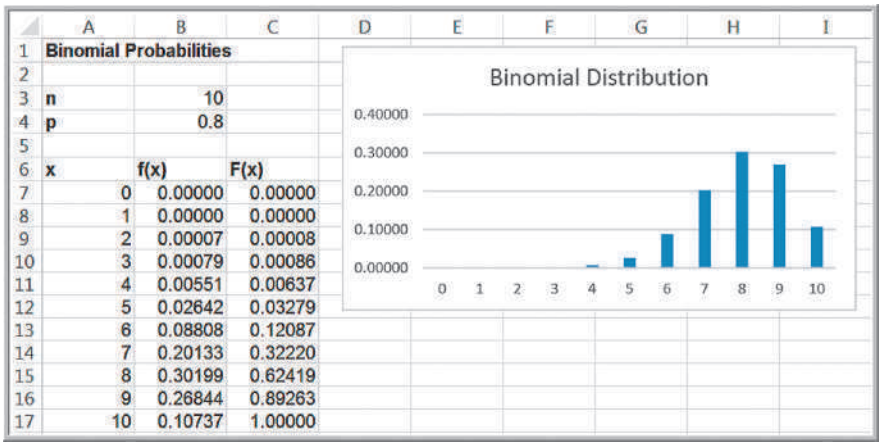
Computing Binomial Probabilities in Excel



	A	B	C
1	**Binomial Probabilities**		
2			
3	**n**	10	
4	**p**	0.2	
5			
6	**x**	**f(x)**	**F(x)**
7	0	0.10737	0.10737
8	1	0.26844	0.37581
9	2	0.30199	0.67780
10	3	0.20133	0.87913
11	4	0.08808	0.96721
12	5	0.02642	0.99363
13	6	0.00551	0.99914
14	7	0.00079	0.99992
15	8	0.00007	1.00000
16	9	0.00000	1.00000
17	10	0.00000	1.00000

Figure 5.11

Example of the Binomial Distribution with $p = 0.8$



	A	B	C
1	**Binomial Probabilities**		
2			
3	n	10	
4	p	0.8	
5			
6	x	f(x)	F(x)
7	0	0.00000	0.00000
8	1	0.00000	0.00000
9	2	0.00007	0.00008
10	3	0.00079	0.00086
11	4	0.00551	0.00637
12	5	0.02642	0.03279
13	6	0.08808	0.12087
14	7	0.20133	0.32220
15	8	0.30199	0.62419
16	9	0.26844	0.89263
17	10	0.10737	1.00000

The expected value of the binomial distribution is np, and the variance is $np(1 - p)$. The binomial distribution can assume different shapes and amounts of skewness, depending on the parameters. Figure 5.11 shows an example when $p = 0.8$. For larger values of p, the binomial distribution is negatively skewed; for smaller values, it is positively skewed. When $p = 0.5$, the distribution is symmetric.

Poisson Distribution

The **Poisson distribution** is a discrete distribution used to model the number of occurrences in some unit of measure—for example, the number of customers arriving at a Subway store during a weekday lunch hour, the number of failures of a machine during a month, number of visits to a Web page during 1 minute, or the number of errors per line of software code.

The Poisson distribution assumes no limit on the number of occurrences (meaning that the random variable X may assume any nonnegative integer value), that occurrences are independent, and that the average number of occurrences per unit is a constant, λ (Greek lowercase lambda). The expected value of the Poisson distribution is λ, and the variance also is equal to λ.

The probability mass function for the Poisson distribution is:

$$f(x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!}, & \text{for } x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases} \tag{5.14}$$

EXAMPLE 5.27 Computing Poisson Probabilities

Suppose that, on average, the number of customers arriving at Subway during lunch hour is 12 customers per hour. The probability that exactly x customers will arrive during the hour is given by a Poisson distribution with a mean of 12. The probability that exactly x customers will arrive during the hour would be calculated using formula (5.14):

$$f(x) = \begin{cases} \dfrac{e^{-12}12^x}{x!}, & \text{for } x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

Substituting $x = 5$ in this formula, the probability that exactly 5 customers will arrive is $f(5) = 0.1274$.

Like the binomial, Poisson probabilities are cumbersome to compute by hand. Probabilities can easily be computed in Excel using the function POISSON.DIST(*x, mean, cumulative*).

EXAMPLE 5.28 Using Excel's Poisson Distribution Function

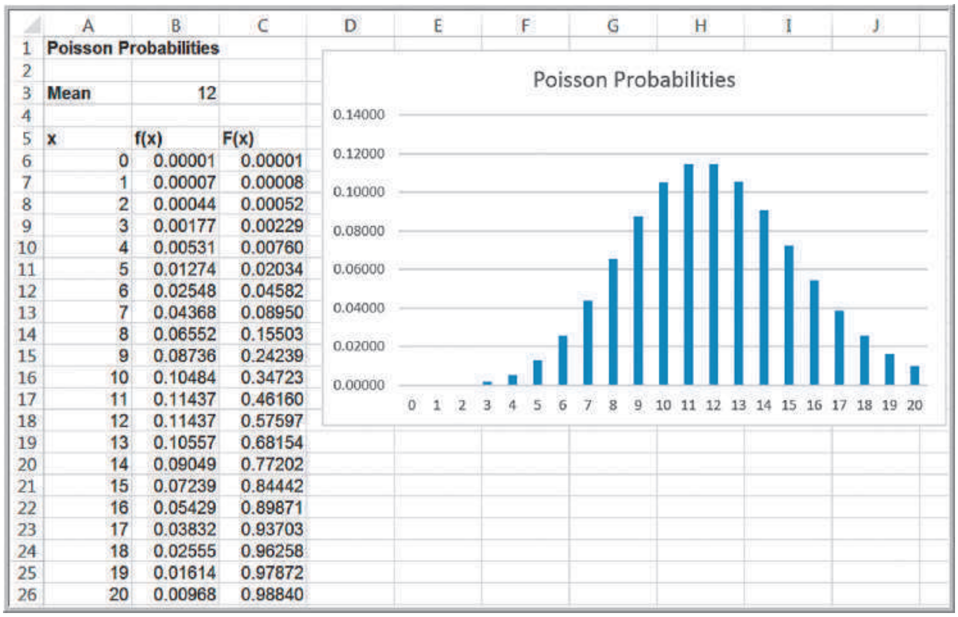
Figure 5.12 shows the results of using this function to compute the distribution for Example 5.26 with $\lambda = 12$ (see the Excel file *Poisson Probabilities*). Thus, the probability of exactly one arrival during the lunch hour is calculated by the Excel function =POISSON.DIST(A7,$B$3,FALSE) = 0.00007 = $f(1)$; the probability of 4 arrivals or fewer is calculated by =POISSON.DIST(A10,$B$3,TRUE) = 0.00760 = $F(4)$, and so on. Because the possible values of a Poisson random variable are infinite, we have not shown the complete distribution. As x gets large, the probabilities become quite small. Like the binomial, the specific shape of the distribution depends on the value of the parameter λ; the distribution is more skewed for smaller values.

Continuous Probability Distributions

As we noted earlier, a continuous random variable is defined over one or more intervals of real numbers and, therefore, has an infinite number of possible outcomes. Suppose that the expert who predicted the probabilities associated with next year's change in the DJIA in Figure 5.6 kept refining the estimates over larger and larger ranges of values. Figure 5.13

Figure 5.12

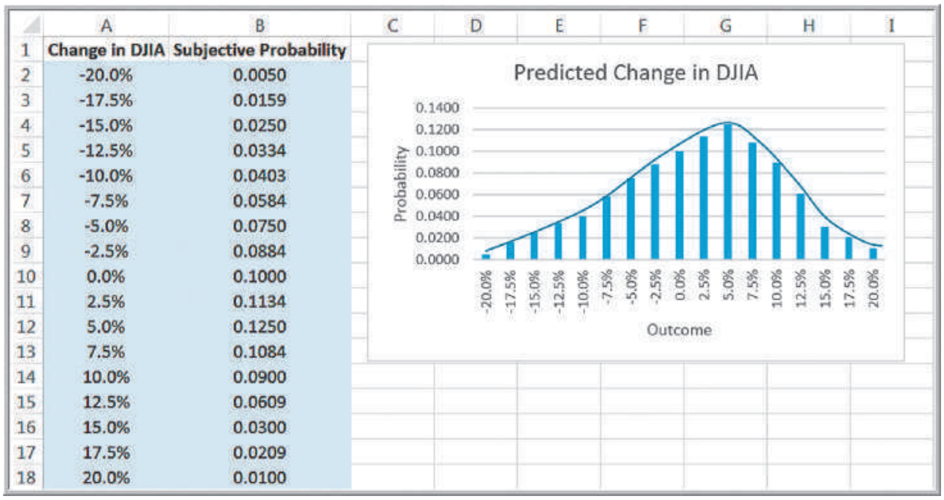
Computing Poisson Probabilities in Excel



	A	B	C
1	**Poisson Probabilities**		
2			
3	**Mean**	12	
4			
5	**x**	**f(x)**	**F(x)**
6	0	0.00001	0.00001
7	1	0.00007	0.00008
8	2	0.00044	0.00052
9	3	0.00177	0.00229
10	4	0.00531	0.00760
11	5	0.01274	0.02034
12	6	0.02548	0.04582
13	7	0.04368	0.08950
14	8	0.06552	0.15503
15	9	0.08736	0.24239
16	10	0.10484	0.34723
17	11	0.11437	0.46160
18	12	0.11437	0.57597
19	13	0.10557	0.68154
20	14	0.09049	0.77202
21	15	0.07239	0.84442
22	16	0.05429	0.89871
23	17	0.03832	0.93703
24	18	0.02555	0.96258
25	19	0.01614	0.97872
26	20	0.00968	0.98840

Figure 5.13

Refined Probability Distribution of DJIA Change



	A	B
1	**Change in DJIA**	**Subjective Probability**
2	-20.0%	0.0050
3	-17.5%	0.0159
4	-15.0%	0.0250
5	-12.5%	0.0334
6	-10.0%	0.0403
7	-7.5%	0.0584
8	-5.0%	0.0750
9	-2.5%	0.0884
10	0.0%	0.1000
11	2.5%	0.1134
12	5.0%	0.1250
13	7.5%	0.1084
14	10.0%	0.0900
15	12.5%	0.0609
16	15.0%	0.0300
17	17.5%	0.0209
18	20.0%	0.0100

Analytics in Practice: Using the Poisson Distribution for Modeling Bids on Priceline[2]

Priceline is well known for allowing customers to name their own prices (but not the service providers) in bidding for services such as airline flights or hotel stays. Some hotels take advantage of Priceline's approach to fill empty rooms for leisure travelers while not diluting the business market by offering discount rates through traditional channels. In one study using business analytics to develop a model to optimize pricing strategies for Kimpton Hotels, which develops, owns, or manages more than 40 independent boutique lifestyle hotels in the United States and Canada, the distribution of the number of bids for a given number of days before arrival was modeled as a Poisson distribution because it corresponded well with data that were observed. For example, the average number of bids placed per day 3 days before arrival on a weekend (the random variable X) was 6.3. Therefore, the distribution used in the model was $f(x) = e^{-6.3}6.3^x/x!$, where x is the number of bids placed. The analytic model helped to determine the prices to post on Priceline and the inventory allocation for each price. After using the model, rooms sold via Priceline increased 11% in 1 year, and the average rate for these rooms increased 3.7%.

Lucas Photo/Shutterstock.com

shows what such a probability distribution might look like using 2.5% increments rather than 5%. Notice that the distribution is similar in shape to the one in Figure 5.6 but simply has more outcomes. If this refinement process continues, then the distribution will approach the shape of a smooth curve, as shown in the figure. Such a curve that characterizes outcomes of a continuous random variable is called a **probability density function** and is described by a mathematical function $f(x)$.

Properties of Probability Density Functions

A probability density function has the following properties:

1. *$f(x) \geq 0$ for all values of x.* This means that a graph of the density function must lie at or above the x-axis.
2. *The total area under the density function above the x-axis is 1.0.* This is analogous to the property that the sum of all probabilities of a discrete random variable must add to 1.0.
3. $P(X = x) = 0$. For continuous random variables, it does not make mathematical sense to attempt to define a probability for a specific value of x because there are an infinite number of values.

[2] Based on Chris K. Anderson, "Setting Prices on Priceline," *Interfaces*, 39, 4 (July–August 2009): 307–315.

4. *Probabilities of continuous random variables are only defined over intervals*. Thus, we may calculate probabilities between two numbers a and b, $P(a \leq X \leq b)$, or to the left or right of a number c—for example, $P(X < c)$ and $P(X > c)$.
5. $P(a \leq X \leq b)$ *is the area under the density function between a and b.*

The cumulative distribution function for a continuous random variable is denoted the same way as for discrete random variables, $F(x)$, and represents the probability that the random variable X is less than or equal to x, $P(X \leq x)$. Intuitively, $F(x)$ represents the area under the density function to the left of x. $F(x)$ can often be derived mathematically from $f(x)$.

Knowing $F(x)$ makes it easy to compute probabilities over intervals for continuous distributions. The probability that X is between a and b is equal to the difference of the cumulative distribution function evaluated at these two points; that is,

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \tag{5.15}$$

For continuous distributions we need not be concerned about the endpoints, as we were with discrete distributions, because $P(a \leq X \leq b)$ is the same as $P(a < X < b)$.

The formal definitions of expected value and variance for a continuous random variable are similar to those for a discrete random variable; however, to understand them, we must rely on notions of calculus, so we do not discuss them in this book. We simply state them when appropriate.

Uniform Distribution

The **uniform distribution** characterizes a continuous random variable for which all outcomes between some minimum and maximum value are equally likely. The uniform distribution is often assumed in business analytics applications when little is known about a random variable other than reasonable estimates for minimum and maximum values. The parameters a and b are chosen judgmentally to reflect a modeler's best guess about the range of the random variable.

For a uniform distribution with a minimum value a and a maximum value b, the density function is

$$f(x) = \begin{cases} \dfrac{1}{b - a}, & \text{for } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{5.16}$$

and the cumulative distribution function is

$$F(x) = \begin{cases} 0, & \text{if } x < a \\ \dfrac{x - a}{b - a}, & \text{if } a \leq x \leq b \\ 1, & \text{if } b < x \end{cases} \tag{5.17}$$

Although Excel does not provide a function to compute uniform probabilities, the formulas are simple enough to incorporate into a spreadsheet. Probabilities are also easy to compute for the uniform distribution because of the simple geometric shape of the density function, as Example 5.29 illustrates.

EXAMPLE 5.29 Computing Uniform Probabilities

Suppose that sales revenue, X, for a product varies uniformly each week between $a = \$1000$ and $b = \$2000$. The density function is $f(x) = 1/(2000 - 1000) = 1/1000$ and is shown in Figure 5.14. Note that the area under the density is function is 1.0, which you can easily verify by multiplying the height by the width of the rectangle.

Suppose we wish to find the probability that sales revenue will be less than $x = \$1{,}300$. We could do this in two ways. First, compute the area under the density function using geometry, as shown in Figure 5.15. The area is $(1/1{,}000)(300) = 0.30$. Alternatively, we could use formula (5.17) to compute $F(1{,}300)$:

$$F(1{,}300) = (1{,}300 - 1{,}000)/(2{,}000 - 1{,}000) = 0.30$$

In either case, the probability is 0.30.

Now suppose we wish to find the probability that revenue will be between \$1,500 and \$1,700. Again, using geometrical arguments (see Figure 5.16), the area of the rectangle between \$1,500 and \$1,700 is $(1/1{,}000)(200) = 0.2$. We may also use formula (5.15) and compute it as follows:

$$\begin{aligned} P(1{,}500 \le X \le 1{,}700) &= P(X \le 1{,}700) - P(X \le 1{,}500) \\ &= F(1{,}700) - F(1{,}500) \\ &= \frac{(1{,}700 - 1{,}000)}{(2{,}000 - 1{,}000)} - \frac{(1{,}500 - 1{,}000)}{(2{,}000 - 1{,}000)} \\ &= 0.7 - 0.5 = 0.2 \end{aligned}$$

The expected value and variance of a uniform random variable X are computed as follows:

$$\mathrm{E}[X] = \frac{a + b}{2} \quad \textbf{(5.18)}$$

$$\mathrm{Var}[X] = \frac{(b - a)^2}{12} \quad \textbf{(5.19)}$$

A variation of the uniform distribution is one for which the random variable is restricted to integer values between a and b (also integers); this is called a **discrete uniform**

Figure 5.14

Uniform Probability Density Function

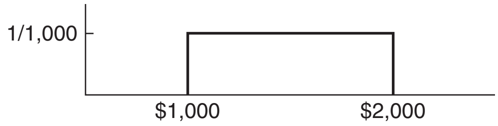


Figure 5.15

Probability that $X < \$1{,}300$

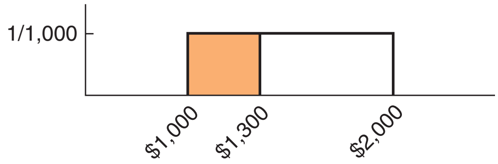
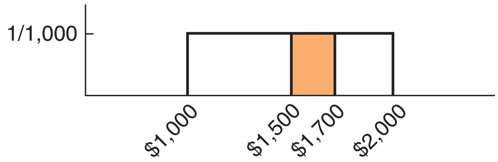


Figure 5.16

$P(\$1{,}500 < X < \$1{,}700)$



distribution. An example of a discrete uniform distribution is the roll of a single die. Each of the numbers 1 through 6 has a $\frac{1}{6}$ probability of occurrence.

Normal Distribution

The **normal distribution** is a continuous distribution that is described by the familiar bell-shaped curve and is perhaps the most important distribution used in statistics. The normal distribution is observed in many natural phenomena. Test scores such as the SAT, deviations from specifications of machined items, human height and weight, and many other measurements are often normally distributed.

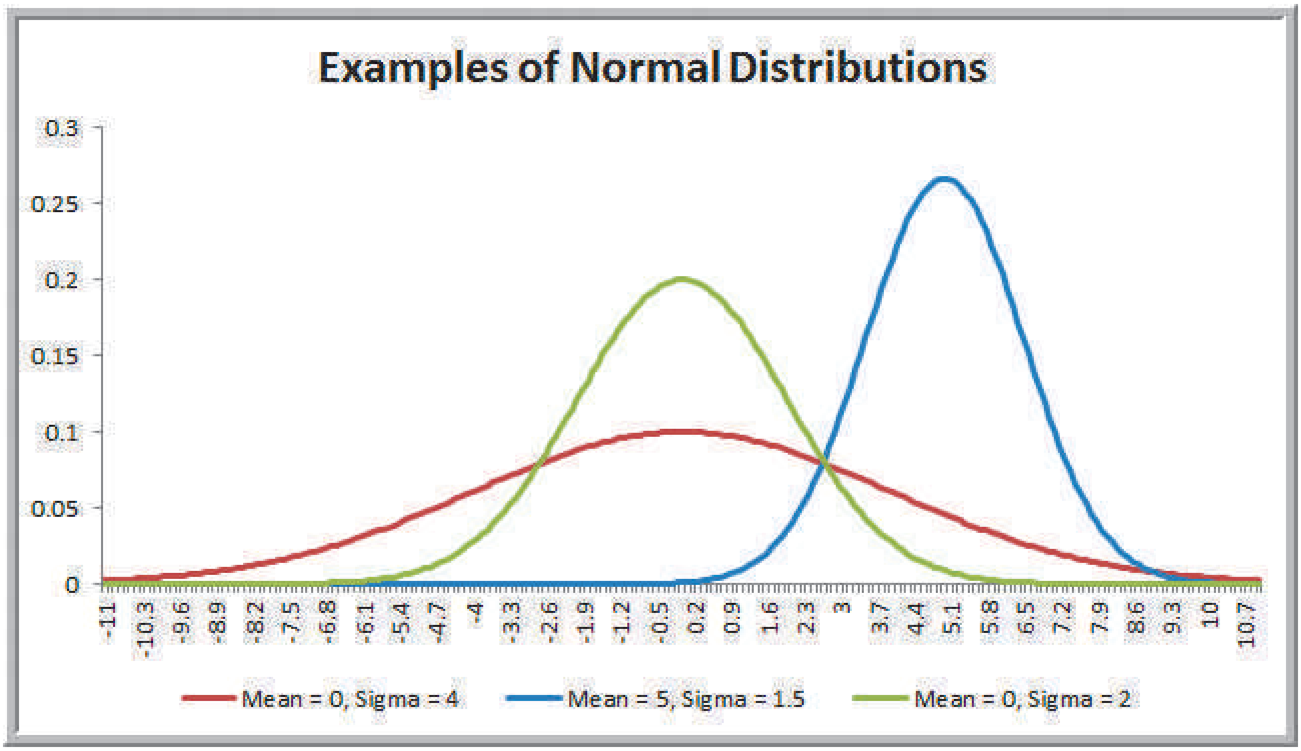
The normal distribution is characterized by two parameters: the mean, μ, and the standard deviation, σ. Thus, as μ changes, the location of the distribution on the x-axis also changes, and as σ is decreased or increased, the distribution becomes narrower or wider, respectively. Figure 5.17 shows some examples.

The normal distribution has the following properties:

1. The distribution is symmetric, so its measure of skewness is zero.
2. The mean, median, and mode are all equal. Thus, half the area falls above the mean and half falls below it.
3. The range of X is unbounded, meaning that the tails of the distribution extend to negative and positive infinity.
4. The empirical rules apply exactly for the normal distribution; the area under the density function within ±1 standard deviation is 68.3%, the area under the density function within ±2 standard deviation is 95.4%, and the area under the density function within ±3 standard deviation is 99.7%.

Figure 5.17

Examples of Normal Distributions



Normal probabilities cannot be computed using a mathematical formula. Instead, we may use the Excel function NORM.DIST(x, *mean*, *standard_deviation*, *cumulative*). NORM.DIST(x, *mean*, *standard_deviation*, *TRUE*) calculates the cumulative probability $F(x) = P(X \leq x)$ for a specified mean and standard deviation. (If *cumulative* is set to *FALSE*, the function simply calculates the value of the density function $f(x)$, which has little practical application other than tabulating values of the density function. This was used to draw the distributions in Figure 5.17.)

EXAMPLE 5.30 Using the NORM.DIST Function to Compute Normal Probabilities

Suppose that a company has determined that the distribution of customer demand (X) is normal with a mean of 750 units/month and a standard deviation of 100 units/month. Figure 5.18 shows some cumulative probabilities calculated with the NORM.DIST function (see the Excel file *Normal Probabilities*). The company would like to know the following:

1. What is the probability that demand will be at most 900 units?
2. What is the probability that demand will exceed 700 units?
3. What is the probability that demand will be between 700 and 900 units?

To answer the questions, first draw a picture. This helps to ensure that you know what area you are trying to calculate and how to use the formulas for working with a cumulative distribution correctly.

Question 1. Figure 5.19(a) shows the probability that demand will be at most 900 units, or $P(X < 900)$. This is simply the cumulative probability for $x = 900$, which can be calculated using the Excel function =NORM.DIST(900,750,100,TRUE) = 0.9332.

Question 2. Figure 5.19(b) shows the probability that demand will exceed 700 units, $P(X > 700)$. Using the principles we have previously discussed, this can be found by subtracting $P(X < 700)$ from 1:

$$P(X > 700) = 1 - P(X < 700) = 1 - F(700)$$
$$= 1 - 0.3085 = 0.6915$$

This can be computed in Excel using the formula =1 – NORM.DIST (700,750,100,TRUE).

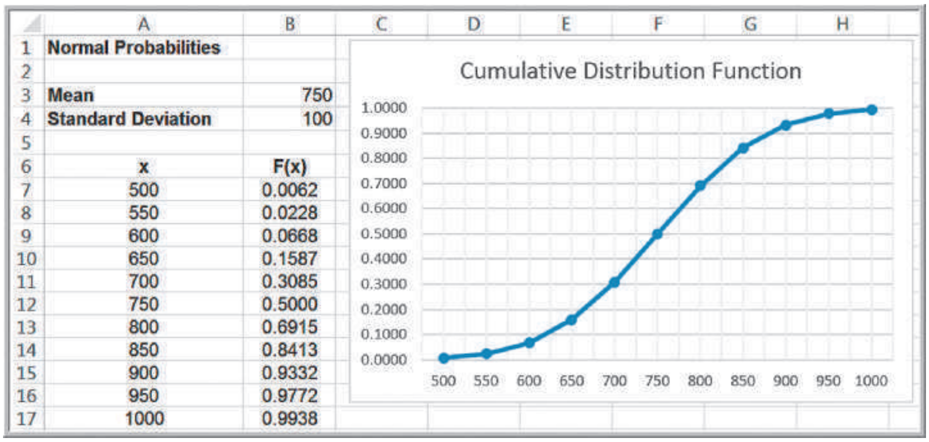
Question 3. The probability that demand will be between 700 and 900, $P(700 < X < 900)$, is illustrated in Figure 5.19(c). This is calculated by

$$P(700 < X < 900) = P(X < 900) - P(X < 700)$$
$$= F(900) - F(700) = 0.9332 - 0.3085 = 0.6247$$

In Excel, we would use the formula =NORM.DIST (900,750,100,TRUE) – NORM.DIST(700,750,100,TRUE).

Figure 5.18

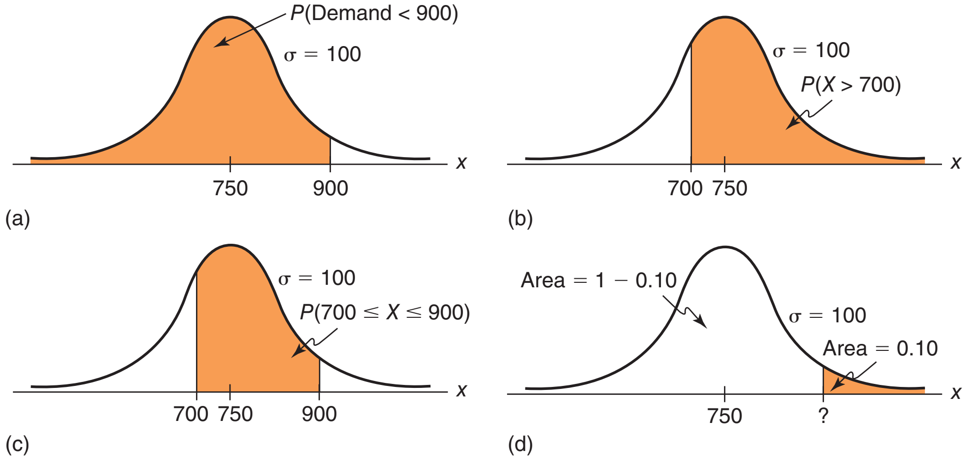
Normal Probability Calculations in Excel



	A	B
1	**Normal Probabilities**	
2		
3	**Mean**	750
4	**Standard Deviation**	100
5		
6	x	F(x)
7	500	0.0062
8	550	0.0228
9	600	0.0668
10	650	0.1587
11	700	0.3085
12	750	0.5000
13	800	0.6915
14	850	0.8413
15	900	0.9332
16	950	0.9772
17	1000	0.9938

Figure 5.19

Computing Normal Probabilities



The NORM.INV Function

With the NORM.DIST function, we are given a value of the random variable X and can find the cumulative probability to the left of x. Now let's reverse the problem. Suppose that we know the cumulative probability but don't know the value of x. How can we find it? We are often faced with such a question in many applications. The Excel function NORM.INV(*probability*, *mean*, *standard_dev*) can be used to do this. In this function, *probability* is the cumulative probability value corresponding to the value of x we seek "INV" stands for inverse.

EXAMPLE 5.31 Using the NORM.INV Function

In the previous example, what level of demand would be exceeded at most 10% of the time? Here, we need to find the value of x so that $P(X > x) = 0.10$. This is illustrated in Figure 5.19(d). Because the area in the upper tail of the normal distribution is 0.10, the cumulative probability must be $1 - 0.10 = 0.90$. From Figure 5.18, we can see that the correct value must be somewhere between 850 and 900 because $F(850) = 0.8413$ and $F(900) = 0.9332$. We can find the exact value using the Excel function = NORM.INV (0.90,750,100) = 878.155, Therefore, a demand of approximately 878 will satisfy the criterion.

Standard Normal Distribution

Figure 5.20 provides a sketch of a special case of the normal distribution called the **standard normal distribution**—the normal distribution with $\mu = 0$ and $\sigma = 1$. This distribution is important in performing many probability calculations. A standard normal random variable is usually denoted by Z, and its density function by $f(z)$. The scale along the z-axis represents the number of standard deviations from the mean of zero. The Excel function NORM.S.DIST(z) finds probabilities for the standard normal distribution.

EXAMPLE 5.32 Computing Probabilities with the Standard Normal Distribution

We have previously noted that the empirical rules apply to any normal distribution. Let us find the areas under the standard normal distribution within 1, 2, and 3 standard deviations of the mean. These can be found by using the function NORM.S.DIST(z). Figure 5.21 shows a tabulation of the cumulative probabilities for z ranging from -3 to $+3$ and calculations of the areas within 1, 2, and 3 standard deviations of the mean. We apply formula (5.15) to find the difference between the cumulative probabilities, $F(b) - F(a)$. For example, the area within 1 standard deviation of the mean is found by calculating $P(-1 < Z < 1) = F(1) - F(-1) =$ NORM.S.DIST(1) $-$ NORM.S.DIST(-1) $= 0.84134 - 0.15866 = 0.6827$ (the difference due to decimal rounding). As the empirical rules stated, about 68% of the area falls within 1 standard deviation; 95%, within 2 standard deviations; and more than 99%, within 3 standard deviations of the mean.

Figure 5.20

Standard Normal Distribution

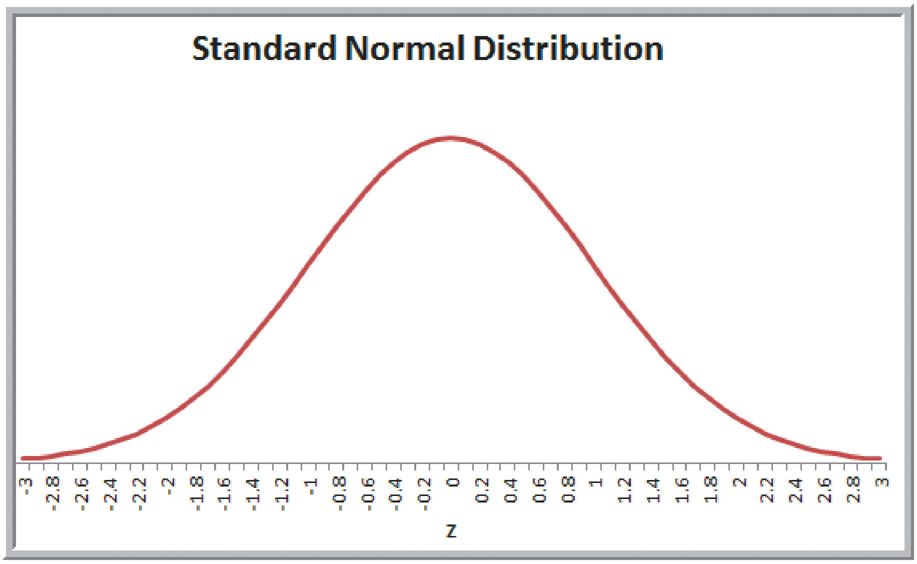
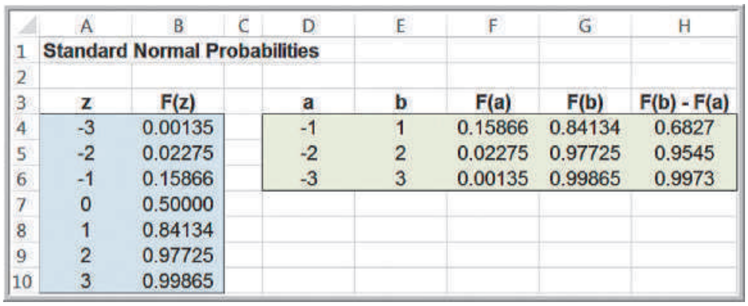


Figure 5.21

Computing Standard Normal Probabilities



	A	B	C	D	E	F	G	H
1	Standard Normal Probabilities							
2								
3	z	F(z)		a	b	F(a)	F(b)	F(b) - F(a)
4	-3	0.00135		-1	1	0.15866	0.84134	0.6827
5	-2	0.02275		-2	2	0.02275	0.97725	0.9545
6	-1	0.15866		-3	3	0.00135	0.99865	0.9973
7	0	0.50000						
8	1	0.84134						
9	2	0.97725						
10	3	0.99865						

Using Standard Normal Distribution Tables

Although it is quite easy to use Excel to compute normal probabilities, tables of the standard normal distribution are commonly found in textbooks and professional references when a computer is not available. Such a table is provided in Table A.1 of Appendix A at the end of this book. The table allows you to look up the cumulative probability for any value of z between -3.00 and $+3.00$.

One of the advantages of the standard normal distribution is that we may compute probabilities for any normal random variable X having a mean μ and standard deviation σ by converting it to a standard normal random variable Z. We introduced the concept of standardized values (z-scores) for sample data in Chapter 4. Here, we use a similar formula to convert a value x from an arbitrary normal distribution into an equivalent standard normal value, z:

$$z = \frac{(x - \mu)}{\sigma} \tag{5.20}$$

EXAMPLE 5.33 Computing Probabilities with Standard Normal Tables

We will answer the first question posed in Example 5.30: What is the probability that demand will be at most $x = 900$ units if the distribution of customer demand (X) is normal with a mean of 750 units/month and a standard deviation of 100 units/month? Using formula (5.19), convert x to a standard normal value:

$$z = \frac{900 - 750}{100} = 1.5$$

Note that 900 is 150 units higher than the mean of 750; since the standard deviation is 100, this simply means that 900 is 1.5 standard deviations above the mean, which is the value of z. Using Table A.1 in Appendix A, we see that the cumulative probability for $z = 1.5$ is 0.9332, which is the same answer we found for Example 5.30.

Exponential Distribution

The **exponential distribution** is a continuous distribution that models the time between randomly occurring events. Thus, it is often used in such applications as modeling the time between customer arrivals to a service system or the time to or between failures of machines, lightbulbs, hard drives, and other mechanical or electrical components.

Similar to the Poisson distribution, the exponential distribution has one parameter, λ. In fact, the exponential distribution is closely related to the Poisson; if the number of events occurring *during* an interval of time has a Poisson distribution, then the time *between* events is exponentially distributed. For instance, if the number of arrivals at a bank is Poisson-distributed, say with mean $\lambda = 12/\text{hour}$ then the time between arrivals is exponential, with mean $\mu = 1/12$ hour, or 5 minutes.

The exponential distribution has the density function

$$f(x) = \lambda e^{-\lambda x}, \quad \text{for } x \geq 0 \tag{5.21}$$

and its cumulative distribution function is

$$F(x) = 1 - e^{-\lambda x}, \quad \text{for } x \geq 0 \tag{5.22}$$

Sometimes, the exponential distribution is expressed in terms of the mean μ rather than the rate λ. To do this, simply substitute $1/\mu$ for λ in the preceding formulas.

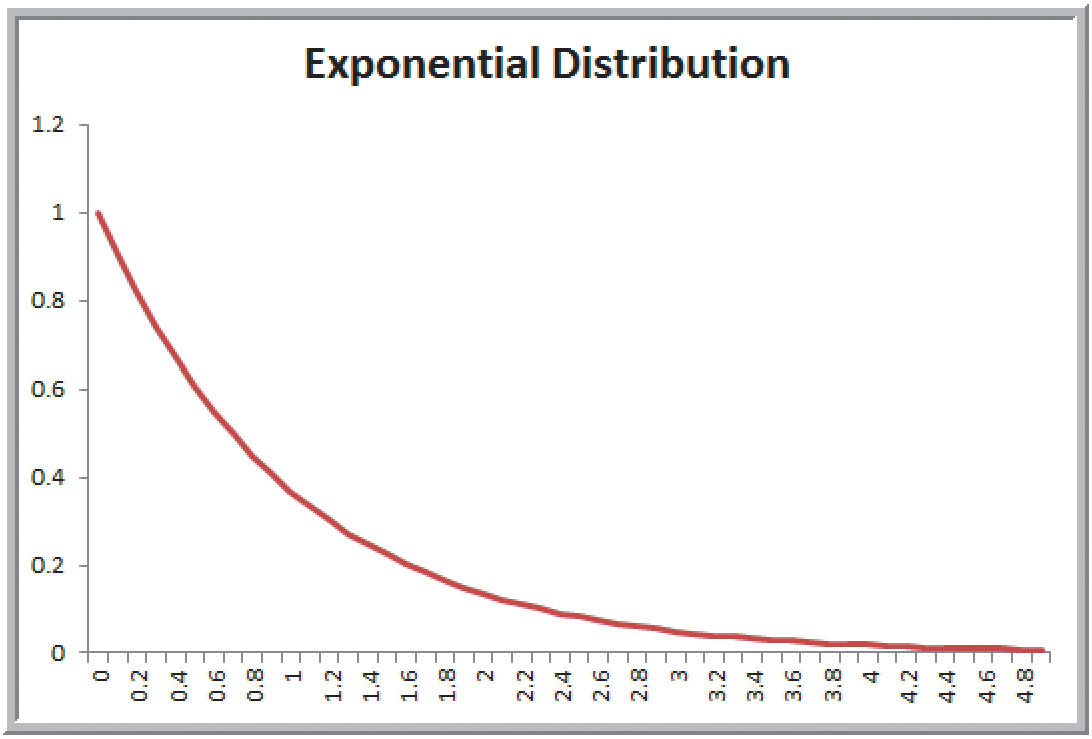
The expected value of the exponential distribution is $1/\lambda$ and the variance is $(1/\lambda)^2$. Figure 5.22 provides a sketch of the exponential distribution. The exponential distribution has the properties that it is bounded below by 0, it has its greatest density at 0, and the density declines as x increases. The Excel function EXPON.DIST (x, *lambda*, *cumulative*) can be used to compute exponential probabilities. As with other Excel probability distribution functions, *cumulative* is either TRUE or FALSE, with TRUE providing the cumulative distribution function.

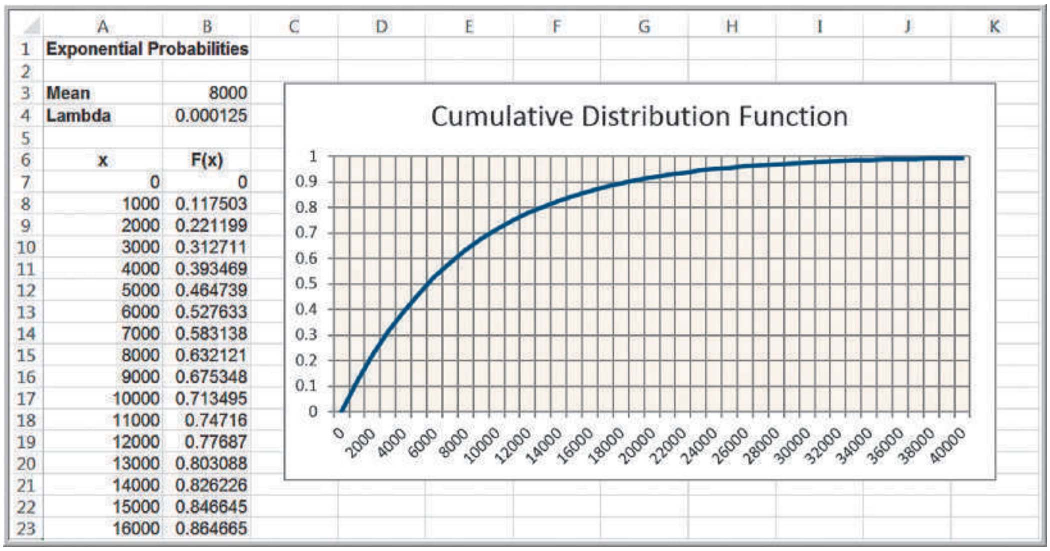
EXAMPLE 5.34 Using the Exponential Distribution

Suppose that the mean time to failure of a critical component of an engine is $\mu = 8{,}000$ hours. Therefore, $\lambda = 1/\mu = 1/8{,}000$ failures/hour. The probability that the component will fail before x hours is given by the cumulative distribution function $F(x)$. Figure 5.23 shows a portion of the cumulative distribution function, which may be found in the Excel file *Exponential Probabilities*. For example, the probability of failing before 5,000 hours is $F(5000) = 0.4647$.

Figure 5.22

Example of an Exponential Distribution ($\lambda = 1$)





	A	B
1	**Exponential Probabilities**	
2		
3	**Mean**	8000
4	**Lambda**	0.000125
5		
6	**x**	**F(x)**
7	0	0
8	1000	0.117503
9	2000	0.221199
10	3000	0.312711
11	4000	0.393469
12	5000	0.464739
13	6000	0.527633
14	7000	0.583138
15	8000	0.632121
16	9000	0.675348
17	10000	0.713495
18	11000	0.74716
19	12000	0.77687
20	13000	0.803088
21	14000	0.826226
22	15000	0.846645
23	16000	0.864665

Figure 5.23

Computing Exponential Probabilities in Excel

Other Useful Distributions

Many other probability distributions, especially those distributions that assume a wide variety of shapes, find application in decision models for characterizing a wide variety of phenomena. Such distributions provide a great amount of flexibility in representing both empirical data or when judgment is needed to define an appropriate distribution. We provide a brief description of these distributions; however, you need not know the mathematical details about them to use them in applications.

Continuous Distributions

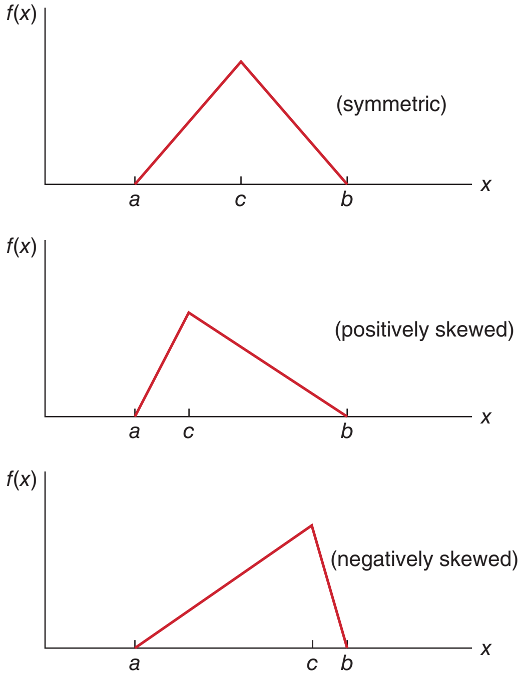
Triangular Distribution. The triangular distribution is defined by three parameters: the minimum, a; maximum, b; and most likely, c. Outcomes near the most likely value have a higher chance of occurring than those at the extremes. By varying the most likely value, the triangular distribution can be symmetric or skewed in either direction, as shown in Figure 5.24. The triangular distribution is often used when no data are available to characterize an uncertain variable and the distribution must be estimated judgmentally.

Lognormal Distribution. If the natural logarithm of a random variable X is normal, then X has a lognormal distribution. Because the lognormal distribution is positively skewed and bounded below by zero, it finds applications in modeling phenomena that have low probabilities of large values and cannot have negative values, such as the time to complete a task. Other common examples include stock prices and real estate prices. The lognormal distribution is also often used for "spiked" service times, that is, when the probability of zero is very low, but the most likely value is just greater than zero.

Beta Distribution. One of the most flexible distributions for modeling variation over a fixed interval from 0 to a positive value is the beta. The beta distribution is a function of two parameters, α and β, both of which must be positive. If α and β are equal, the distribution is symmetric. If either parameter is 1.0 and the other is greater than 1.0, the distribution is in the shape of a J. If α is

Figure 5.24

Examples of Triangular Distributions

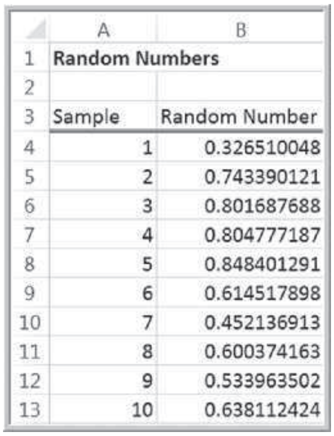


less than β, the distribution is positively skewed; otherwise, it is negatively skewed. These properties can help you to select appropriate values for the shape parameters.

Random Sampling from Probability Distributions

Many applications in business analytics require random samples from specific probability distributions. For example, in a financial model, we might be interested in the distribution of the cumulative discounted cash flow over several years when sales, sales growth rate, operating expenses, and inflation factors are all uncertain and are described by probability distributions. The outcome variables of such decision models are complicated functions of the random input variables. Understanding the probability distribution of such variables can be accomplished only by sampling procedures called Monte Carlo simulation, which we address in Chapter 12.

The basis for generating random samples from probability distributions is the concept of a random number. A **random number** is one that is uniformly distributed between 0 and 1. Technically speaking, computers cannot generate truly random numbers since they must use a predictable algorithm. However, the algorithms are designed to generate a sequence of numbers that appear to be random. In Excel, we may generate a random number within any cell using the function RAND( ). This function has no arguments; therefore, nothing should be placed within the parentheses (but the parentheses are required). Figure 5.25 shows a table of 10 random numbers generated in Excel. You should be aware that unless the automatic recalculation feature is suppressed, whenever any cell in the spreadsheet is modified, the values in any cell containing the RAND( ) function will change. Automatic recalculation can be changed to manual by choosing *Calculation Options* in the *Calculation* group under the *Formulas* tab. Under manual recalculation mode, the worksheet is recalculated only when the F9 key is pressed.



	A	B
1	Random Numbers	
2		
3	Sample	Random Number
4	1	0.326510048
5	2	0.743390121
6	3	0.801687688
7	4	0.804777187
8	5	0.848401291
9	6	0.614517898
10	7	0.452136913
11	8	0.600374163
12	9	0.533963502
13	10	0.638112424

Figure 5.25

A Sample of Random Numbers

Sampling from Discrete Probability Distributions

Sampling from discrete probability distributions using random numbers is quite easy. We will illustrate this process using the probability distribution for rolling two dice.

EXAMPLE 5.35 Sampling from the Distribution of Dice Outcomes

The probability mass function and cumulative distribution in decimal form are as follows:

x	$f(x)$	$F(x)$
2	0.0278	0.0278
3	0.0556	0.0833
4	0.0833	0.1667
5	0.1111	0.2778
6	0.1389	0.4167
7	0.1667	0.5833
8	0.1389	0.7222
9	0.1111	0.8333
10	0.0833	0.9167
11	0.0556	0.9722
12	0.0278	1.0000

Notice that the values of $F(x)$ divide the interval from 0 to 1 into smaller intervals that correspond to the probabilities of the outcomes. For example, the interval from (but not including) 0 and up to and including 0.0278 has a probability of 0.028 and corresponds to the outcome $x = 2$; the interval from (but not including) 0.0278 and up to and including 0.0833 has a probability of 0.0556 and corresponds to the outcome $x = 3$; and so on. This is summarized as follows:

Interval	Outcome
0 to 0.0278	2
0.0278 to 0.0833	3
0.0833 to 0.1667	4
0.1667 to 0.2778	5
0.2778 to 0.4167	6
0.4167 to 0.5833	7
0.5833 to 0.7222	8
0.7222 to 0.8323	9
0.8323 to 0.9167	10
0.9167 to 0.9722	11
0.9722 to 1.0000	12

Any random number, then, must fall within one of these intervals. Thus, to generate an outcome from this distribution, all we need to do is to select a random number and determine the interval into which it falls. Suppose we use the data in Figure 5.25. The first random number is 0.326510048. This falls in the interval corresponding to the sample outcome of 6. The second random number is 0.743390121. This number falls in the interval corresponding to an outcome of 9. Essentially, we have developed a technique to roll dice on a computer. If this is done repeatedly, the frequency of occurrence of each outcome should be proportional to the size of the random number range (i.e., the probability associated with the outcome) because random numbers are uniformly distributed.

We can easily use this approach to generate outcomes from any discrete distribution; the VLOOKUP function in Excel can be used to implement this on a spreadsheet.

EXAMPLE 5.36 Using the VLOOKUP Function for Random Sampling

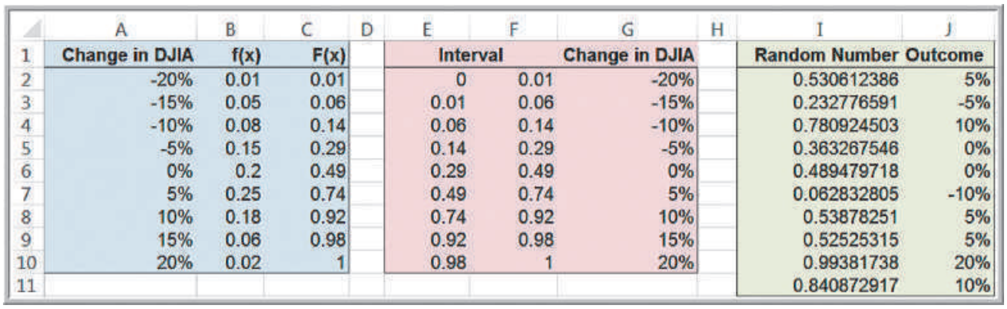
Suppose that we want to sample from the probability distribution of the predicted change in the Dow Jones Industrial Average index shown in Figure 5.6. We first construct the cumulative distribution *F*(*x*). Then assign intervals to the outcomes based on the values of the cumulative distribution, as shown in Figure 5.26. This specifies the table range for the VLOOKUP function, namely, $E$2:$G$10. List the random numbers in a column using the RAND( ) function. The formula in cell J2 is =VLOOKUP(I2,$E$2:$G$10,3), which is copied down that column. This function takes the value of the random number in cell I2, finds the last number in the first column of the table range that is less than the random number, and returns the value in the third column of the table range. In this case, 0.49 is the last number in column E that is less than 0.530612386, so the function returns 5% as the outcome.

Sampling from Common Probability Distributions

This approach of generating random numbers and transforming them into outcomes from a probability distribution may be used to sample from most any distribution. A value randomly generated from a specified probability distribution is called a **random variate**. For example, it is quite easy to transform a random number into a random variate from a uniform distribution between *a* and *b*. Consider the formula:

$$U = a + (b - a)*\text{RAND}(\,) \qquad \textbf{(5.23)}$$

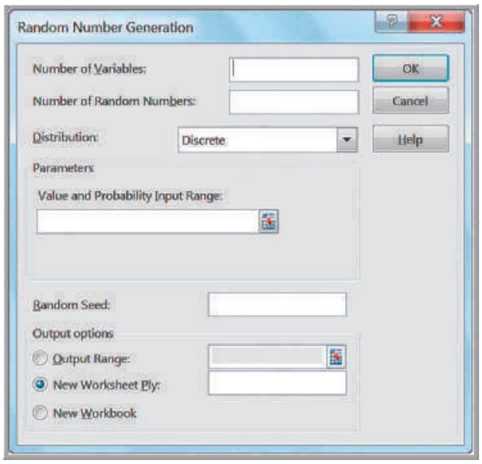
Note that when RAND( ) $= 0$, $U = a$, and when RAND( ) approaches 1, U approaches b. For any other value of RAND( ) between 0 and 1, $(b - a)$*RAND( ) represents the same proportion of the interval (a, b) as RAND( ) does of the interval $(0, 1)$. Thus, all



	A	B	C	D	E	F	G	H	I	J
1	Change in DJIA	f(x)	F(x)		Interval		Change in DJIA		Random Number	Outcome
2	-20%	0.01	0.01		0	0.01	-20%		0.530612386	5%
3	-15%	0.05	0.06		0.01	0.06	-15%		0.232776591	-5%
4	-10%	0.08	0.14		0.06	0.14	-10%		0.780924503	10%
5	-5%	0.15	0.29		0.14	0.29	-5%		0.363267546	0%
6	0%	0.2	0.49		0.29	0.49	0%		0.489479718	0%
7	5%	0.25	0.74		0.49	0.74	5%		0.062832805	-10%
8	10%	0.18	0.92		0.74	0.92	10%		0.53878251	5%
9	15%	0.06	0.98		0.92	0.98	15%		0.52525315	5%
10	20%	0.02	1		0.98	1	20%		0.99381738	20%
11									0.840872917	10%

Figure 5.26

Using the VLOOKUP Function to Sample from a Discrete Distribution



Distribution: Discrete

● New Worksheet Ply:

Figure 5.27

Excel *Random Number Generation* Dialog

real numbers between *a* and *b* can occur. Since RAND( ) is uniformly distributed, so also is *U*.

Although this is quite easy, it is certainly not obvious how to generate random variates from other distributions such as normal or exponential. We do not describe the technical details of how this is done but rather just describe the capabilities available in Excel. Excel allows you to generate random variates from discrete distributions and certain others using the *Random Number Generation* option in the *Analysis Toolpak*. From the *Data* tab in the ribbon, select *Data Analysis* in the *Analysis* group and then *Random Number Generation*. The *Random Number Generation* dialog, shown in Figure 5.27, will appear. From the *Random Number Generation* dialog, you may select from seven distributions: uniform, normal, Bernoulli, binomial, Poisson, and patterned, as well as discrete. (The patterned distribution is characterized by a lower and upper bound, a step, a repetition rate for values, and a repetition rate for the sequence.) If you select the *Output Range* option, you are asked to specify the upper-left cell reference of the output table that will store the outcomes, the number of variables (columns of values you want generated), number of random numbers (the number of data points you want generated for each variable), and the type of distribution. The default distribution is the discrete distribution.

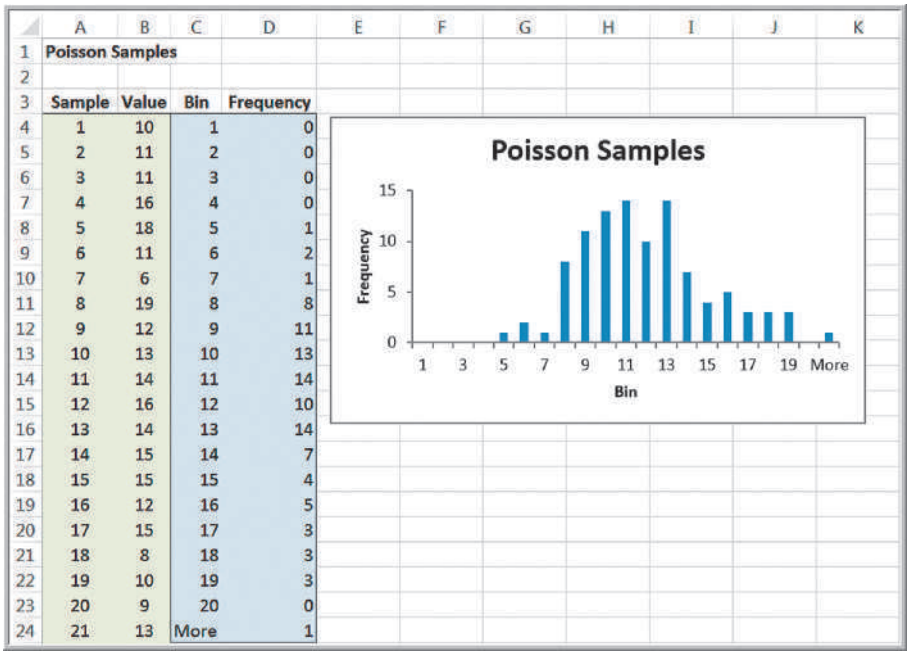
EXAMPLE 5.37 Using Excel's *Random Number Generation* Tool

We will generate 100 outcomes from a Poisson distribution with a mean of 12. In the *Random Number Generation* dialog, set the *Number of Variables* to 1 and the *Number of Random Numbers* to 100 and select Poisson from the drop-down *Distribution* box. The dialog will change and prompt you for the value of *Lambda*, the mean of the Poisson distribution; enter 12 in the box and click *OK*. The tool will display the random numbers in a column. Figure 5.28 shows a histogram of the results.

The dialog in Figure 5.27 also allows you the option of specifying a random number seed. A **random number seed** is a value from which a stream of random numbers

Figure 5.28

Histogram of Samples from a Poisson Distribution



	A	B	C	D
1	Poisson Samples			
2				
3	Sample	Value	Bin	Frequency
4	1	10	1	0
5	2	11	2	0
6	3	11	3	0
7	4	16	4	0
8	5	18	5	1
9	6	11	6	2
10	7	6	7	1
11	8	19	8	8
12	9	12	9	11
13	10	13	10	13
14	11	14	11	14
15	12	16	12	10
16	13	14	13	14
17	14	15	14	7
18	15	15	15	4
19	16	12	16	5
20	17	15	17	3
21	18	8	18	3
22	19	10	19	3
23	20	9	20	0
24	21	13	More	1

is generated. By specifying the same seed, you can produce the same random numbers at a later time. This is desirable when we wish to reproduce an identical sequence of "random" events in a simulation to test the effects of different policies or decision variables under the same circumstances. However, one disadvantage with using the *Random Number Generation* tool is that you must repeat the process to generate a new set of sample values; pressing the recalculation (F9) key will not change the values. This can make it difficult to use this tool to analyze decision models.

Excel also has several inverse functions of probability distributions that may be used to generate random variates. For the normal distribution, use

- NORM.INV(*probability*, *mean*, *standard_deviation*)—normal distribution with a specified mean and standard deviation,
- NORM.S.INV(*probability*)—standard normal distribution.

For some advanced distributions, you might see

- LOGNORM.INV(*probability*, *mean*, *standard_deviation*)—lognormal distribution, where ln(*X*) has the specified mean and standard deviation,
- BETA.INV(*probability*, *alpha*, *beta*, *A*, *B*)—beta distribution.

To use these functions, simply enter RAND( ) in place of *probability* in the function. For example, NORM.INV(RAND( ), 5, 2) will generate random variates from a normal distribution with mean 5 and standard deviation 2. Each time the worksheet is recalculated, a new random number and, hence, a new random variate, are generated. These functions may be embedded in cell formulas and will generate new values whenever the worksheet is recalculated.

The following example shows how sampling from probability distributions can provide insights about business decisions that would be difficult to analyze mathematically.

EXAMPLE 5.38 A Sampling Experiment for Evaluating Capital Budgeting Projects

In finance, one way of evaluating capital budgeting projects is to compute a profitability index (*PI*), which is defined as the ratio of the present value of future cash flows (*PV*) to the initial investment (*I*):

$$PI = PV/I \tag{5.24}$$

Because the cash flow and initial investment that may be required for a particular project are often uncertain, the profitability index is also uncertain. If we can characterize *PV* and *I* by some probability distributions, then we would like to know the probability distribution for *PI*. For example, suppose that *PV* is estimated to be normally distributed with a mean of $12 million and a standard deviation of $2.5 million, and the initial investment is also estimated to be normal with a mean of $3.0 million and standard deviation of $0.8 million. Intuitively, we might believe that the profitability index is also normally distributed with a mean of $12 million/$3 million = $4 million; however, as we shall see, this is not the case. We can use a sampling experiment to identify the probability distribution of *PI* for these assumptions.

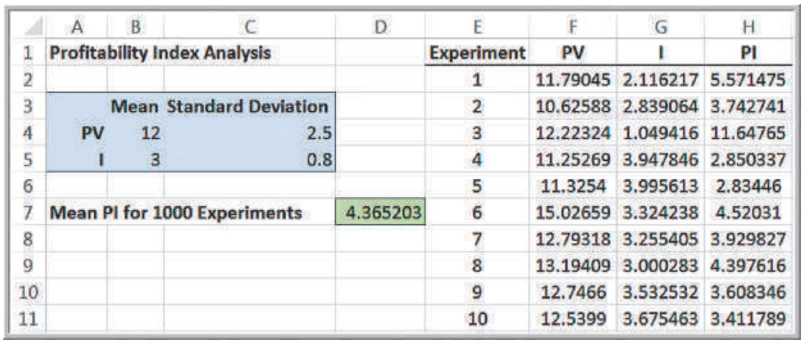
Figure 5.29 shows a simple model from the Excel file *Profitability Index Experiment*. For each experiment, the values of *PV* and *I* are sampled from their assumed normal distributions using the NORM.INV function. *PI* is calculated in column D, and the average value for 1,000 experiments is shown in cell E8. We clearly see that this is not equal to 4 as previously suspected. The histogram in Figure 5.30 also demonstrates that the distribution of *PI* is not normal but is skewed to the right. This experiment confirms that the ratio of two normal distributions is not normally distributed. We encourage you to create this spreadsheet and replicate this experiment (note that your results will not be exactly the same as these because you are generating random values!)

Probability Distribution Functions in *Analytic Solver Platform*

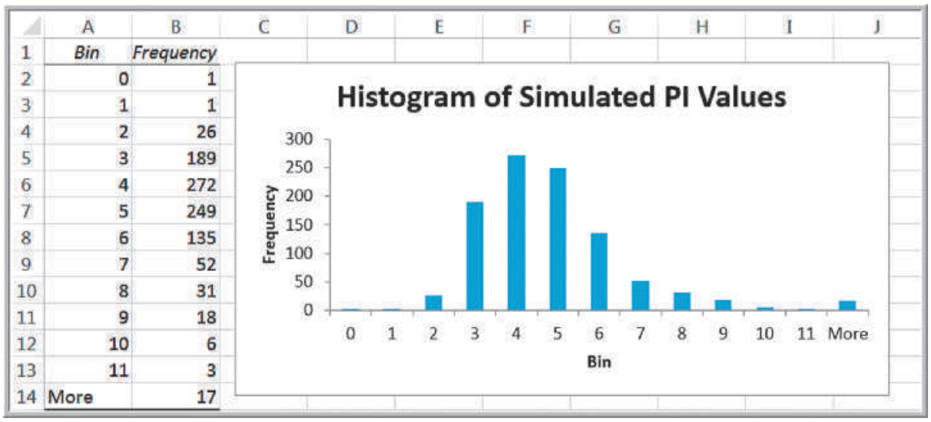
Analytic Solver Platform (see the section on Spreadsheet Add-ins in Chapter 2) provides custom Excel functions that generate random samples from specified probability distributions. Table 5.1 shows a list of these for distributions we have discussed. These functions return random values from the specified distributions in worksheet cells. These functions will be very useful in business analytics applications in later chapters, especially Chapter 12 on simulation and risk analysis.

Figure 5.29

Sampling Experiment for Profitability Index



	A	B	C	D	E	F	G	H
1	Profitability Index Analysis				Experiment	PV	I	PI
2					1	11.79045	2.116217	5.571475
3		Mean	Standard Deviation		2	10.62588	2.839064	3.742741
4	PV	12	2.5		3	12.22324	1.049416	11.64765
5	I	3	0.8		4	11.25269	3.947846	2.850337
6					5	11.3254	3.995613	2.83446
7	Mean PI for 1000 Experiments			4.365203	6	15.02659	3.324238	4.52031
8					7	12.79318	3.255405	3.929827
9					8	13.19409	3.000283	4.397616
10					9	12.7466	3.532532	3.608346
11					10	12.5399	3.675463	3.411789



Bin	Frequency
0	1
1	1
2	26
3	189
4	272
5	249
6	135
7	52
8	31
9	18
10	6
11	3
More	17

Figure 5.30

Frequency Distribution and Histogram of Profitability Index

Table 5.1

Analytic Solver Platform Probability Distribution Functions

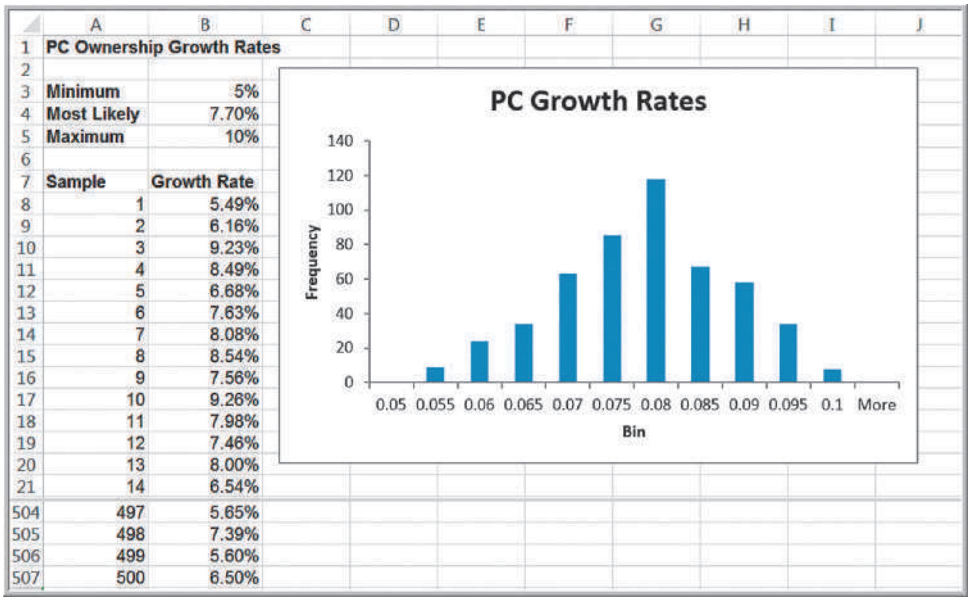
Distribution	*Analytic Solver Platform* Function
Bernoulli	PsiBernoulli(*probability*)
Binomial	PsiBinomial(*trials*, *probability*)
Poisson	PsiPoisson(*mean*)
Uniform	PsiUniform(*lower*, *upper*)
Normal	PsiNormal(*mean*, *standard deviation*)
Exponential	PsiExponential(*mean*)
Discrete Uniform	PsiDisUniform(*values*)
Geometric	PsiGeometric(*probability*)
Negative Binomial	PsiNegBinomial(*successes*, *probability*)
Hypergeometric	PsiHyperGeo(*trials*, *success*, *population size*)
Triangular	PsiTriangular(*minimum*, *most likely*, *maximum*)
Lognormal	PsiLognormal(*mean*, *standard deviation*)
Beta	PsiBeta(*alpha*, *beta*)

EXAMPLE 5.39 Using *Analytic Solver Platform* Distribution Functions

An energy company was considering offering a new product and needed to estimate the growth in PC ownership. Using the best data and information available, they determined that the minimum growth rate was 5.0%, the most likely value was 7.7%, and the maximum value was 10.0%. These parameters characterize a triangular distribution. Figure 5.31 (Excel file *PC Ownership Growth Rates*) shows a portion of 500 samples that were generated using the function PsiTriangular(5%, 7.7%, 10%). Notice that the histogram exhibits a clear triangular shape.

Figure 5.31

Samples from a Triangular Distribution



	A	B
1	PC Ownership Growth Rates	
2		
3	Minimum	5%
4	Most Likely	7.70%
5	Maximum	10%
6		
7	Sample	Growth Rate
8	1	5.49%
9	2	6.16%
10	3	9.23%
11	4	8.49%
12	5	6.68%
13	6	7.63%
14	7	8.08%
15	8	8.54%
16	9	7.56%
17	10	9.26%
18	11	7.98%
19	12	7.46%
20	13	8.00%
21	14	6.54%
504	497	5.65%
505	498	7.39%
506	499	5.60%
507	500	6.50%

Data Modeling and Distribution Fitting

In many applications of business analytics, we need to collect sample data of important variables such as customer demand, purchase behavior, machine failure times, and service activity times, to name just a few, to gain an understanding of the distributions of these variables. Using the tools we have studied, we may construct frequency distributions and histograms and compute basic descriptive statistical measures to better understand the nature of the data. However, sample data are just that—samples.

Using sample data may limit our ability to predict uncertain events that may occur because potential values *outside* the range of the sample data are not included. A better approach is to identify the underlying probability distribution from which sample data come by "fitting" a theoretical distribution to the data and verifying the goodness of fit statistically.

To select an appropriate theoretical distribution that fits sample data, we might begin by examining a histogram of the data to look for the distinctive shapes of particular distributions. For example, normal data are symmetric, with a peak in the middle. Exponential data are very positively skewed, with no negative values. Lognormal data are also very positively skewed, but the density drops to zero at 0. Various forms of the gamma, Weibull, or beta distributions could be used for distributions that do not seem to fit one of the other common forms. This approach is not, of course, always accurate or valid, and sometimes it can be difficult to apply, especially if sample sizes are small. However, it may narrow the search down to a few potential distributions.

Summary statistics can also provide clues about the nature of a distribution. The mean, median, standard deviation, and coefficient of variation often provide information about the nature of the distribution. For instance, normally distributed data tend to have a fairly low coefficient of variation (however, this may not be true if the mean is small). For normally distributed data, we would also expect the median and mean to be approximately the same. For exponentially distributed data, however, the median will be less than the mean. Also, we would expect the mean to be about equal to the standard deviation, or, equivalently, the coefficient of variation would be close to 1. We could also look at the skewness index. Normal data are not skewed, whereas lognormal and exponential data are positively skewed. The following examples illustrate some of these ideas.

EXAMPLE 5.40 Analyzing Airline Passenger Data

An airline operates a daily route between two medium-sized cities using a 70-seat regional jet. The flight is rarely booked to capacity but often accommodates business travelers who book at the last minute at a high price. Figure 5.32 shows the number of passengers for a sample of 25 flights (Excel file *Airline Passengers*). The histogram shows a relatively symmetric distribution. The mean, median, and mode are all similar, although there is some degree of positive skewness. From our discussion in Chapter 4 about the variability of samples, it is important to recognize that this is a relatively small sample that can exhibit a lot of variability compared with the population from which it is drawn. Thus, based on these characteristics, it would not be unreasonable to assume a normal distribution for the purpose of developing a predictive or prescriptive analytics model.

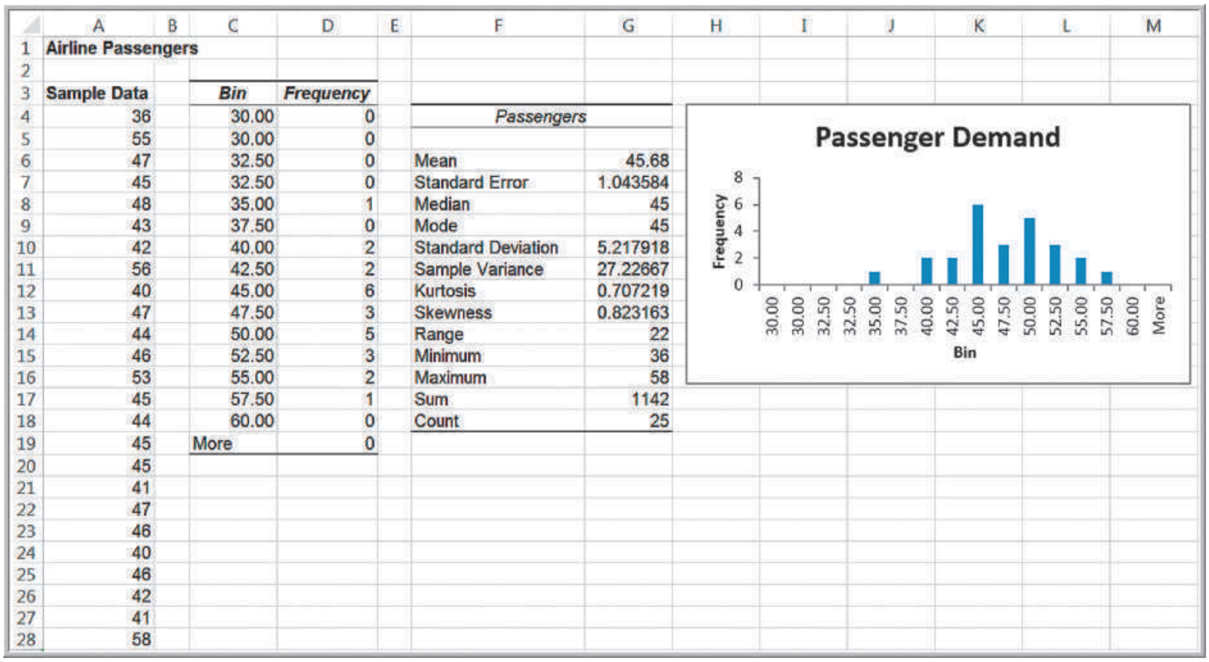
EXAMPLE 5.41 Analyzing Airport Service Times

Figure 5.33 shows a portion of the data and statistical analysis of 812 samples of service times at an airport's ticketing counter (Excel file *Airport Service Times*). It is not clear what the distribution might be. It does not appear to be exponential, but it might be lognormal or even another distribution with which you might not be familiar. From the descriptive statistics, we can see that the mean is not close to the standard deviation, suggesting that the data are probably not exponential. The data are positively skewed, suggesting that a lognormal distribution might be appropriate. However, it is difficult to make a definitive conclusion.

The examination of histograms and summary statistics might provide some idea of the appropriate distribution; however, a better approach is to analytically fit the data to the best type of probability distribution.

Figure 5.32

Data and Statistics for Passenger Demand

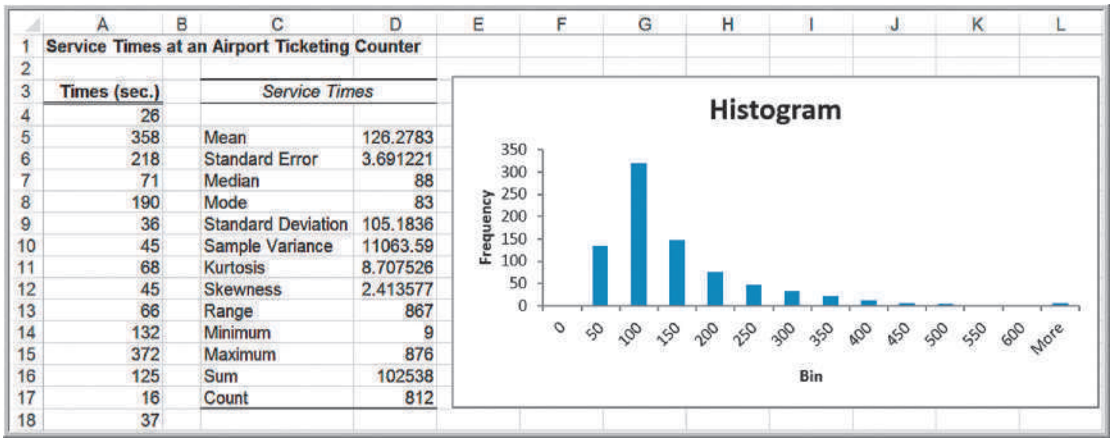


Airline Passengers

Sample Data
36
55
47
45
48
43
42
56
40
47
44
46
53
45
44
45
45
41
47
46
40
46
42
41
58

Bin	*Frequency*
30.00	0
30.00	0
32.50	0
32.50	0
35.00	1
37.50	0
40.00	2
42.50	2
45.00	6
47.50	3
50.00	5
52.50	3
55.00	2
57.50	1
60.00	0
More	0

Passengers	
Mean	45.68
Standard Error	1.043584
Median	45
Mode	45
Standard Deviation	5.217918
Sample Variance	27.22667
Kurtosis	0.707219
Skewness	0.823163
Range	22
Minimum	36
Maximum	58
Sum	1142
Count	25



	A	B	C	D
1	Service Times at an Airport Ticketing Counter			
2				
3	Times (sec.)		*Service Times*	
4	26			
5	358		Mean	126.2783
6	218		Standard Error	3.691221
7	71		Median	88
8	190		Mode	83
9	36		Standard Deviation	105.1836
10	45		Sample Variance	11063.59
11	68		Kurtosis	8.707526
12	45		Skewness	2.413577
13	66		Range	867
14	132		Minimum	9
15	372		Maximum	876
16	125		Sum	102538
17	16		Count	812
18	37			

Figure 5.33

Airport Service Times Statistics

Goodness of Fit

The basis for fitting data to a probability distribution is a statistical procedure called **goodness of fit**. Goodness of fit attempts to draw a conclusion about the *nature* of the distribution. For instance, in Example 5.40 we suggested that it might be reasonable to assume that the distribution of passenger demand is normal. Goodness of fit would provide objective, analytical support for this assumption. Understanding the details of this procedure requires concepts that we will learn in Chapter 7. However, software exists (which we illustrate shortly) that run statistical procedures to determine how well a theoretical distribution fits a set of data, and also find the best-fitting distribution.

Determining how well sample data fits a distribution is typically measured using one of three types of statistics, called chi-square, Kolmogorov-Smirnov, and Anderson-Darling statistics. Essentially, these statistics provide a measure of how well the histogram of the sample data compares with a specified theoretical probability distribution. The chi-square approach breaks down the theoretical distribution into areas of equal probability and compares the data points within each area to the number that would be expected for that distribution. The Kolmogorov-Smirnov procedure compares the cumulative distribution of the data with the theoretical distribution and bases its conclusion on the largest vertical distance between them. The Anderson-Darling method is similar but puts more weight on the differences between the tails of the distributions. This approach is useful when you need a better fit at the extreme tails of the distribution. If you use chi-square, you should have at least 50 data points; for small samples, the Kolmogorov-Smirnov test generally works better.

Distribution Fitting with *Analytic Solver Platform*

Analytic Solver Platform has the capability of "fitting" a probability distribution to data using one of the three goodness-of-fit procedures. This is often done to analyze and define inputs to simulation models that we discuss in Chapter 12. However, you need not understand simulation at this time to use this capability. We illustrate this procedure using the airport service time data.

EXAMPLE 5.42 Fitting a Distribution to Airport Service Times

Step 1: Highlight the range of the data in the *Airport Service Times* worksheet. Click on the *Tools* button in the *Analytic Solver Platform* ribbon and then click *Fit*. This displays the *Fit Options* dialog shown in Figure 5.34.

Step 2: In the *Fit Options* dialog, choose whether to fit the data to a continuous or discrete distribution. In this example, we select *Continuous*. You may also choose the statistical procedure used to evaluate the results, either chi-square, Kolmogorov-Smirnov, or Anderson-Darling. We choose the default option, Kolmogorov-Smirnov. Click the *Fit* button.

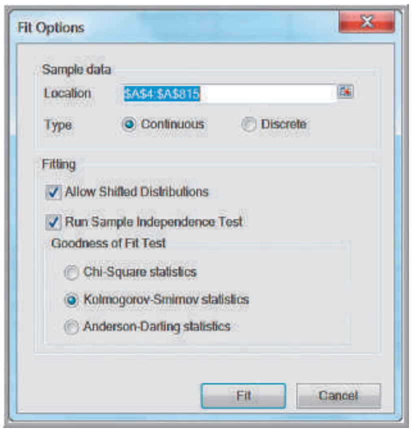
Analytic Solver Platform displays a window with the results as shown in Figure 5.35. In this case, the best-fitting distribution is called an Erlang distribution. If you want to compare the results to a different distribution, simply check the box on the left side. You don't have to know the mathematical details to use the distribution in a spreadsheet application because the formula for the Psi function corresponding to this distribution is shown in the panel on the right side of the output. When you exit the dialog, you have the option to accept the result; if so, it asks you to select a cell to place the Psi function for the distribution, in this case, the function:

=PsiErlang(1.46504838280818,80.0576462180289,
PsiShift 8.99)

We could use this function to generate samples from this distribution, similar to the way we used the NORM.INV function in Example 5.38.

Figure 5.34

Fit Options Dialog



Fit Options

Sample data

Location $A$4:$A$815

Type ● Continuous ○ Discrete

Fitting

☑ Allow Shifted Distributions

☑ Run Sample Independence Test

Goodness of Fit Test

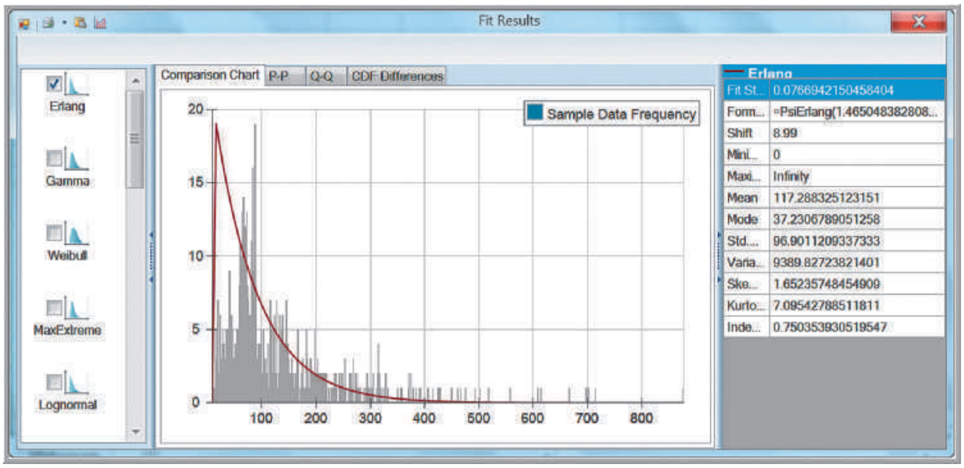
○ Chi-Square statistics

● Kolmogorov-Smirnov statistics

○ Anderson-Darling statistics

Figure 5.35

Analytic Solver Platform Distribution Fitting Results



Erlang

Fit St...	0.0766942150458404
Form...	=PsiErlang(1.465048382808...
Shift	8.99
Mini...	0
Maxi...	Infinity
Mean	117.288325123151
Mode	37.2306789051258
Std....	96.9011209337333
Varia...	9389.82723821401
Ske...	1.65235748454909
Kurto...	7.09542788511811
Inde...	0.750353930519547

Analytics in Practice: The Value of Good Data Modeling in Advertising

To illustrate the importance of identifying the correct distribution in decision modeling, we discuss an example in advertising.[3] The amount that companies spend on the creative component of advertising (i.e., making better ads) is traditionally quite small relative to the overall media budget. One expert noted that the expenditure on creative development was about 5% of that spent on the media delivery campaign.

Whatever money is spent on creative development is usually directed through a single advertising agency. However, one theory that has been proposed is that more should be spent on creative ad development, and the expenditures should be spread across a number of competitive advertising agencies. In research studies of this theory, the distribution of advertising effectiveness was assumed to be normal. In reality, data collected on the response to consumer product ads show that this distribution is actually quite skewed and, therefore, not normally distributed. Using the wrong assumption in any model or application can produce erroneous results. In this situation, the skewness actually provides an advantage for advertisers, making it more effective to obtain ideas from a variety of advertising agencies.

A mathematical model (called Gross's model) relates the relative contributions of creative and media dollars to total advertising effectiveness and is often used to identify the best number of draft ads to purchase. This model includes factors of ad development cost, total media spending budget, the distribution of effectiveness across ads (assumed to be normal), and the unreliability of identifying the most effective ad from a set of independently generated alternatives. Gross's model concluded that large gains were possible if multiple ads were obtained from independent sources, and the best ad is selected.

Victor Correia/Shutterstock.com

Since the data observed on ad effectiveness was clearly skewed, other researchers examined ad effectiveness by studying standard industry data on ad recall without requiring the assumption of normally distributed effects. This analysis found that the best of a number of ads was more effective than any single ad. Further analysis revealed that the optimal number of ads to commission can vary significantly, depending on the shape of the distribution of effectiveness for a single ad.

The researchers developed an alternative to Gross's model. From their analyses, they found that as the number of draft ads was increased, the effectiveness of the best ad also increased. Both the optimal number of draft ads and the payoff from creating multiple independent drafts were higher *when the correct distribution was used* than the results reported in Gross's original study.

Key Terms

Bernoulli distribution
Binomial distribution
Complement
Conditional probability
Continuous random variable
Cumulative distribution function
Discrete random variable
Discrete uniform distribution

[3] Based on G. C. O'Connor, T. R. Willemain, and J. MacLachlan, "The Value of Competition Among Agencies in Developing Ad Campaigns: Revisiting Gross's Model," *Journal of Advertising*, 25, 1 (1996): 51–62.

Empirical probability distribution
Event
Expected value
Experiment
Exponential distribution
Goodness of fit
Independent events
Intersection
Joint probability
Joint probability table
Marginal probability
Multiplication law of probability
Mutually exclusive
Normal distribution
Outcome
Poisson distribution
Probability
Probability density function
Probability distribution
Probability mass function
Random number
Random number seed
Random variable
Random variate
Sample space
Standard normal distribution
Uniform distribution
Union

Problems and Exercises

1. a. A die is rolled. Find the probability that the number obtained is greater than 4.

   b. Two coins are tossed. Find the probability that only one head is obtained.

   c. Two dice are rolled. Find the probability that the sum is equal to 5.

   d. A card is drawn at random from a deck of cards. Find the probability of getting the King of Hearts.

2. Consider the experiment of drawing two cards without replacement from a deck consisting of only the ace through 10 of a single suit (e.g., only hearts).

   a. Describe the outcomes of this experiment. List the elements of the sample space.

   b. Define the event A_i to be the set of outcomes for which the sum of the values of the cards is i (with an ace $=$ 1). List the outcomes associated with A_i for $i =$ 3 to 19.

   c. What is the probability of obtaining a sum of the two cards equaling from 3 to 19?

3. Find the probability of getting the each of the total values when two dice is rolled: 1, 2, 5, 6, 7, 10, and 11.

4. The students of a class have elected five candidates to represent them on the college management council:

S.No.	Gender	Age
1	Male	18
2	Male	19
3	Female	22
4	Female	20
5	Male	23

This group decides to elect a spokesperson by randomly drawing a name from a hat. Calculate the probability of the spokesperson being either female or over 21.

5. Refer to the card scenario described in Problem 2.

   a. Let A be the event "total card value is odd." Find $P(A)$ and $P(A^c)$.

   b. What is the probability that the sum of the two cards will be more than 14?

6. The latest nationwide political poll in a particular country indicates that the probability for the candidate to be a republican is 0.55, a communist is 0.30, and a supporter of the patriots of that country is 0.15. Assuming that these probabilities are accurate, within a randomly chosen group of 10 citizens:

   a. What is the probability that four are communists?

   b. What is the probability that none are republican?

7. Roulette is played at a table similar to the one in Figure 5.36. A wheel with the numbers 1 through 36 (evenly distributed with the colors red and black) and two green numbers 0 and 00 rotates in a shallow bowl with a curved wall. A small ball is spun on the inside of the wall and drops into a pocket corresponding to one of the numbers. Players may make 11 different types of bets by placing chips on different areas of the table. These include bets on a single number, two adjacent numbers, a row of three numbers, a block of four numbers, two adjacent rows of six numbers, and the five number combinations of 0, 00, 1, 2, and 3; bets on the numbers 1–18 or 19–36; the first, second, or third group of 12 numbers; a column of

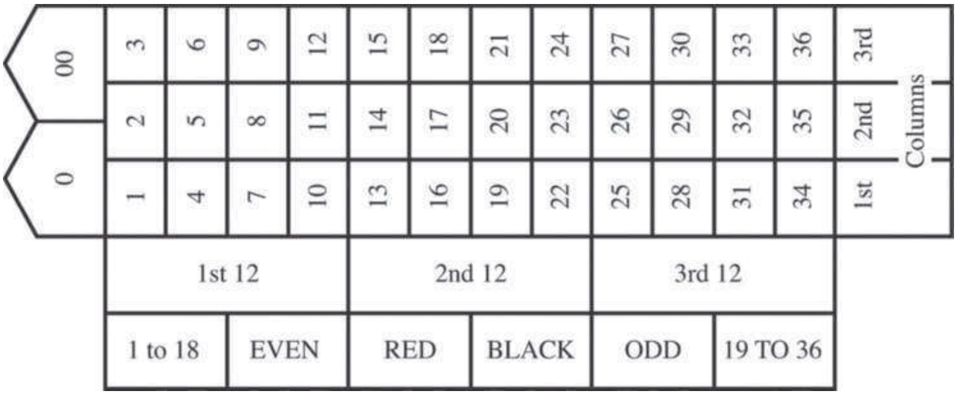


Figure 5.36

Layout of a Typical Roulette Table

12 numbers; even or odd; and red or black. Payoffs differ by bet. For instance, a single-number bet pays 35 to 1 if it wins; a three-number bet pays 11 to 1; a column bet pays 2 to 1; and a color bet pays even money. Define the following events: $C1 =$ column 1 number, $C2 =$ column 2 number, $C3 =$ column 3 number, $O =$ odd number, $E =$ even number, $G =$ green number, $F12 =$ first 12 numbers, $S12 =$ second 12 numbers, and $T12 =$ third 12 numbers.

a. Find the probability of each of these events.

b. Find $P(G \text{ or } O)$, $P(O \text{ or } F12)$, $P(C1 \text{ or } C3)$, $P(E \text{ and } F12)$, $P(E \text{ or } F12)$, $P(S12 \text{ and } T12)$, $P(O \text{ or } C2)$.

8. From a bag full of colored balls (red, blue, green and orange), some are picked out and replaced. This is done a thousand times and the number of times each colored ball is picked out is—Blue: 300, Red: 200, Green: 450, and Orange: 50.

a. What is the probability of picking a green ball?

b. What is the probability of picking a blue ball?

c. If there are 100 balls in the bag, how many of them are likely to be green?

d. If there are 10000 balls in the bag, how many of them are likely to be orange?

9. A box contains marbles of three different colors: 8 black, 6 white, and 4 red. Three marbles are selected at random without replacement. Find the probability that the selection contains each of the outcomes listed.

a. Three black marbles

b. A red, a black and a white marble, in that order

c. A red marble and two white marbles, in any order

10. A survey of 200 college graduates who have been working for at least 3 years found that 90 owned only mutual funds, 20 owned only stocks, and 70 owned both.

a. What is the probability that an individual owns a stock? A mutual fund?

b. What is the probability that an individual owns neither stocks nor mutual funds?

c. What is the probability that an individual owns either a stock or a mutual fund?

11. Row 26 of the Excel file *Census Education Data* gives the number of employed persons having a specific educational level.

a. Find the probability that an employed person has attained each of the educational levels listed in the data.

b. Suppose that A is the event "has at least an Associate's Degree" and B is the event "is at least a high school graduate." Find the probabilities of these events. Are they mutually exclusive? Why or why not? Find the probability $P(A \text{ or } B)$.

12. A survey of shopping habits found the percentage of respondents that use technology for shopping as shown in Figure 5.37. For example, 17.39% only use online coupons; 21.74% use online coupons and check prices online before shopping, and so on.

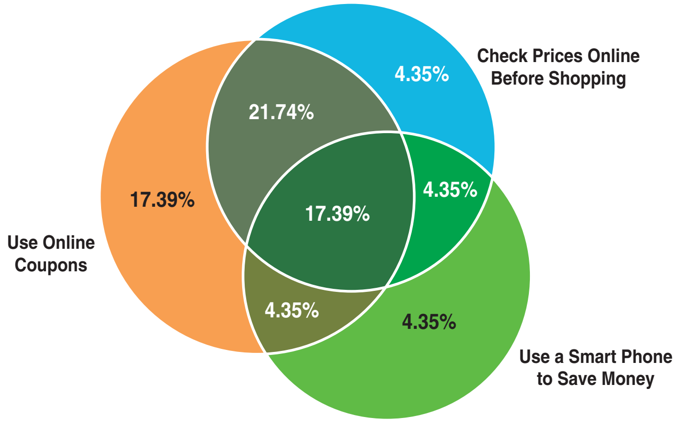
a. What is the probability that a shopper will check prices online before shopping?

b. What is the probability that a shopper will use a smart phone to save money?

c. What is the probability that a shopper will use online coupons?

d. What is the probability that a shopper will not use any of these technologies?

Figure 5.37


Check Prices Online Before Shopping
Use Online Coupons
Use a Smart Phone to Save Money
4.35%
21.74%
17.39%
17.39%
4.35%
4.35%
4.35%

e. What is the probability that a shopper will check prices online and use online coupons but not use a smart phone?

f. If a shopper checks prices online, what is the probability that he or she will use a smart phone?

g. What is the probability that a shopper will check prices online but not use online coupons or a smart phone?

13. A Canadian business school summarized the gender and residency of its incoming class as follows:

	Residency				
Gender	Canada	United States	Europe	Asia	Other
Male	123	24	17	52	8
Female	86	8	10	73	4

a. Construct the joint probability table.

b. Calculate the marginal probabilities.

c. What is the probability that a female student is from outside Canada or the United States?

14. In an example in Chapter 3, we developed the following cross-tabulation of sales transaction data:

Region	Book	DVD	Total
East	56	42	98
North	43	42	85
South	62	37	99
West	100	90	190
Total	261	211	472

a. Find the marginal probabilities that a sale originated in each of the four regions and the marginal probability of each type of sale (book or DVD).

b. Find the conditional probabilities of selling a book given that the customer resides in each region.

15. Use the Civilian Labor Force data in the Excel file *Census Education Data* to find the following:

a. P(unemployed and advanced degree)

b. P(unemployed | advanced degree)

c. P(not a high school grad | unemployed)

d. Are the events "unemployed" and "at least a high school graduate" independent?

16. Using the data in the Excel file *Consumer Transportation Survey*, develop a contingency table for Gender and Vehicle Driven; then convert this table into probabilities.

a. What is the probability that respondent is female?

b. What is the probability that a respondent drives an SUV?

c. What is the probability that a respondent is male and drives a minivan?

d. What is the probability that a female respondent drives either a truck or an SUV?

e. If it is known that an individual drives a car, what is the probability that the individual is female?

f. If it is known that an individual is male, what is the probability that he drives an SUV?

g. Determine whether the random variables "gender" and the event "vehicle driven" are statistically independent. What would this mean for advertisers?

17. A home pregnancy test is not always accurate. Suppose the probability is 0.015 that the test indicates that a woman is pregnant when she actually is not, and the probability is 0.025 that the test indicates that a woman is not pregnant when she really is. Assume that the probability that a woman who takes the test is actually pregnant is 0.7. What is the probability that a woman is pregnant if the test yields a not-pregnant result?

18. A political candidate running for local office is considering the votes she can get in an upcoming election. Assume that the votes can take on only four possible values. If the candidate assessment is per the given Excel sheet *Votes*, construct the probability distribution graph.

Number of Votes	Probability this Will Happen
1000	0.2
2000	0.4
3000	0.3
4000	0.1

19. In the roulette example described in Problem 7, what is the probability that the outcome will be green twice in a row? What is the probability that the outcome will be black twice in a row?

20. A consumer products company found that 48% of successful products also received favorable results from test market research, whereas 12% had unfavorable results but nevertheless were successful. They also found that 28% of unsuccessful products had unfavorable research results, whereas 12% of them had favorable research results. That is, P(successful product and favorable test market) = 0.48, P(successful product and unfavorable test market) = 0.12, P(unsuccessful product and favorable test market) = 0.12, and P(unsuccessful product and unfavorable test market) = 0.28. Find the probabilities of successful and unsuccessful products given known test market results.

21. A particular training program has been designed to upgrade the administrative skills of managers. The program is self-administered; the manager requires putting in different number of hours to complete the program. The previous participant's input indicates that the mean length of time spent on the program is 500 hours, and that this normally distributed random variables has standard deviation of 100 hours. Calculate the probability of a randomly selected participant who will require more than 500 hours.

22. The weekly demand of a slow-moving product has the following probability mass function:

Demand, x	Probability, $f(x)$
0	0.2
1	0.4
2	0.3
3	0.1
4 or more	0

Find the expected value, variance, and standard deviation of weekly demand.

23. The Excel sheet *Baseball* contains information about a team which is using an automatic pitching machine. If the machine is correctly setup and properly adjusted, it will strike 85 percent of the time. If it is incorrectly set up, it will strike only 35 percent of the time. Past data indicates that 75 percent of the setup of the machine is correctly done. After the machine has been set up, at batting practice one day, it throws three strikes on the first three pitches. What is the revised probability that has setup done correctly?

Event	P(Event)	P(1Strike/Event)
Correct	0.75	0.85
Incorrect	x	0.35

24. Based on the data in the Excel file *Consumer Transportation Survey,* develop a probability mass function and cumulative distribution function (both tabular and as charts) for the random variable Number of Children. What is the probability that an individual in this survey has fewer than three children? At least one child? Five or more children?

25. A major application of analytics in marketing is determining the attrition of customers. Suppose that the probability of a long-distance carrier's customer leaving for another carrier from one month to the next is 0.12. What distribution models the retention of an individual customer? What is the expected value and standard deviation?

26. The Excel file *Call Center Data* shows that in a sample of 70 individuals, 27 had prior call center experience. If we assume that the probability that any potential hire will also have experience with a probability of 27/70, what is the probability that among 10 potential hires, more than half of them will have experience? Define the parameter(s) for this distribution based on the data.

27. If a cell phone company conducted a telemarketing campaign to generate new clients and the probability of successfully gaining a new customer was 0.07, what is the probability that contacting 50 potential customers would result in at least 5 new customers?

28. During 1 year, a particular mutual fund has outperformed the S&P 500 index 33 out of 52 weeks. Find the probability that this performance or better would happen again.

29. A popular resort hotel has 300 rooms and is usually fully booked. About 6% of the time a reservation is canceled before the 6:00 p.m. deadline with no penalty. What is the probability that at least 280 rooms will be occupied? Use the binomial distribution to find the exact value.

30. A telephone call center where people place marketing calls to customers has a probability of success of 0.08. The manager is very harsh on those who do not get a sufficient number of successful calls. Find the number of calls needed to ensure that there is a probability of 0.90 of obtaining 5 or more successful calls.

31. Ravi sells three life insurance policies on an average per week. Use Poisson's distribution to calculate the probability that in a given week he will sell

a. some policies.

b. two or more policies but less than 5 policies.

c. one policy, assuming that there are 5 working days per week.

32. The number and frequency of Atlantic hurricanes annually from 1940 through 2012 is shown here.

Number	Frequency
0	5
1	16
2	19
3	14
4	3
5	5
6	4
7	3
8	2
10	1
12	1

a. Find the probabilities of 0–12 hurricanes each season using these data.

b. Assuming a Poisson distribution and using the mean number of hurricanes per season from the empirical data, compute the probabilities of experiencing 0–12 hurricanes in a season. Compare these to your answer to part (a). How good does a Poisson distribution model this phenomenon? Construct a chart to visualize these results.

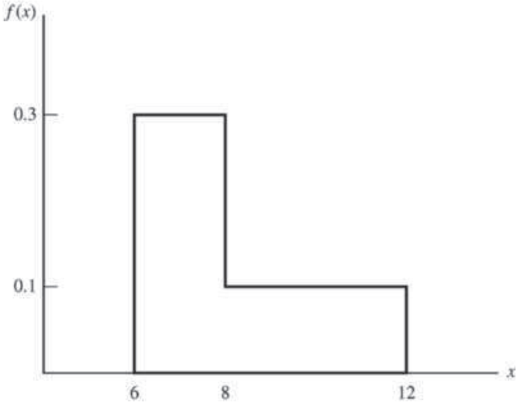
33. Verify that the function corresponding to the following figure is a valid probability density function. Then find the following probabilities:

a. $P(x < 8)$

b. $P(x > 7)$

c. $P(6 < x < 10)$

d. $P(8 < x < 11)$



34. The time required to play a game of Battleship™ is uniformly distributed between 15 and 60 minutes.

a. Find the expected value and variance of the time to complete the game.

b. What is the probability of finishing within 30 minutes?

c. What is the probability that the game would take longer than 40 minutes?

35. A contractor has estimated that the minimum number of days to remodel a bathroom for a client is 10 days. He also estimates that 80% of similar jobs are completed within 18 days. If the remodeling time is uniformly distributed, what should be the parameters of the uniform distribution?

36. In determining automobile-mileage ratings, it was found that the mpg (X) for a certain model is normally distributed, with a mean of 33 mpg and a standard deviation of 1.7 mpg. Find the following:

a. $P(X < 30)$

b. $P(28 < X < 32)$

c. $P(X > 35)$

d. $P(X > 31)$

e. The mileage rating that the upper 5% of cars achieve.

37. The distribution of the SAT scores in math for an incoming class of business students has a mean of 590 and standard deviation of 22. Assume that the scores are normally distributed.

a. Find the probability that an individual's SAT score is less than 550.

b. Find the probability that an individual's SAT score is between 550 and 600.

c. Find the probability that an individual's SAT score is greater than 620.

d. What percentage of students will have scored better than 700?

e. Find the standardized values for students scoring 550, 600, 650, and 700 on the test.

38. A popular soft drink is sold in 2-liter (2,000-milliliter) bottles. Because of variation in the filling process, bottles have a mean of 2,000 milliliters and a standard deviation of 20, normally distributed.

a. If the process fills the bottle by more than 50 milliliters, the overflow will cause a machine malfunction. What is the probability of this occurring?

b. What is the probability of underfilling the bottles by at least 30 milliliters?

39. A supplier contract calls for a key dimension of a part to be between 1.96 and 2.04 centimeters. The supplier has determined that the standard deviation of its process, which is normally distributed, is 0.04 centimeter.

a. If the actual mean of the process is 1.98, what fraction of parts will meet specifications?

b. If the mean is adjusted to 2.00, what fraction of parts will meet specifications?

c. How small must the standard deviation be to ensure that no more than 2% of parts are nonconforming, assuming the mean is 2.00?

40. Dev scored 940 on a national mathematics test. The mean test score was 850 with a standard deviation of 100. What proportion of students had a higher score than Dev? (Assume that the test scores are normally distributed.)

41. A lightbulb is warranted to last for 5,000 hours. If the time to failure is exponentially distributed with a true mean of 4,750 hours, what is the probability that it will last at least 5,000 hours?

42. The actual delivery time from Giodanni's Pizza is exponentially distributed with a mean of 20 minutes.

a. What is the probability that the delivery time will exceed 30 minutes?

b. What proportion of deliveries will be completed within 20 minutes?

43. Develop a procedure to sample from the probability distribution of soft-drink choices in Problem 1. Implement your procedure on a spreadsheet and use the VLOOKUP function to sample 10 outcomes from the distribution.

44. Develop a procedure to sample from the probability distribution of two-card hands in Problem 2. Implement your procedure on a spreadsheet and use the VLOOKUP function to sample 20 outcomes from the distribution.

45. Use formula (5.23) to obtain a sample of 25 outcomes for a game of Battleship™ as described in Problem 34. Find the average and standard deviation for these 25 outcomes.

46. Use the Excel *Random Number Generation* tool to generate 100 samples of the number of customers that the financial consultant in Problem 31 will have on a daily basis. What percentage will meet his target of at least 5?

47. A formula in financial analysis is: Return on equity = net profit margin × total asset turnover × equity multiplier. Suppose that the equity multiplier is fixed at 4.0, but that the net profit margin is normally distributed with a mean of 3.8% and a standard deviation of 0.4%, and that the total asset turnover is normally distributed with a mean of 1.5 and a standard deviation of 0.2. Set up and conduct a sampling experiment to find the distribution of the return on equity. Show your results as a histogram to help explain your analysis and conclusions. Use the empirical rules to predict the return on equity.

48. A government agency is putting a large project out for low bid. Bids are expected from 10 different contractors and will have a normal distribution with a mean of $3.5 million and a standard deviation of $0.25 million. Devise and implement a sampling

experiment for estimating the distribution of the minimum bid and the expected value of the minimum bid.

49. Use *Analytic Solver Platform* to fit the hurricane data in Problem 32 to a discrete distribution? Does the Poisson distribution give the best fit?

50. Use *Analytic Solver Platform* to fit a distribution to the data in the Excel file *Computer Repair Times*. Try the three different statistical measures for evaluating goodness of fit and see if they result in different best-fitting distributions.

51. The Excel file *Investment Returns* provides sample data for the annual return of the S&P 500, and monthly returns of a stock portfolio and bond portfolio. Construct histograms for each data set and use *Analytic Solver Platform* to find the best fitting distribution.

Case: Performance Lawn Equipment

PLE collects a variety of data from special studies, many of which are related to the quality of its products. The company collects data about functional test performance of its mowers after assembly; results from the past 30 days are given in the worksheet *Mower Test*. In addition, many in-process measurements are taken to ensure that manufacturing processes remain in control and can produce according to design specifications. The worksheet *Blade Weight* shows 350 measurements of blade weights taken from the manufacturing process that produces mower blades during the most recent shift. Elizabeth Burke has asked you to study these data from an analytics perspective. Drawing upon your experience, you have developed a number of questions.

1. For the mower test data, what distribution might be appropriate to model the failure of an individual mower?
2. What fraction of mowers fails the functional performance test using all the mower test data?
3. What is the probability of having *x* failures in the next 100 mowers tested, for *x* from 0 to 20?
4. What is the average blade weight and how much variability is occurring in the measurements of blade weights?
5. Assuming that the data are normal, what is the probability that blade weights from this process will exceed 5.20?
6. What is the probability that weights will be less than 4.80?
7. What is the actual percent of weights that exceed 5.20 or are less than 4.80 from the data in the worksheet?
8. Is the process that makes the blades stable over time? That is, are there any apparent changes in the pattern of the blade weights?
9. Could any of the blade weights be considered outliers, which might indicate a problem with the manufacturing process or materials?
10. Was the assumption that blade weights are normally distributed justified? What is the best-fitting probability distribution for the data?

Summarize all your findings to these questions in a well-written report.