GLOBAL EDITION

# Business Analytics

THIRD EDITION

James R. Evans

# Business Analytics

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

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{5.6}$$

## EXAMPLE 5.14 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 (5.6). 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.

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

The conditional probability formula may be used in other ways. For example, multiplying both sides of formula (5.6) 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 $P(B$ and $A)$ is the same as $P(A$ and $B)$; thus we can express $P(A$ and $B)$ in two ways:

$$P(A \text{ and } B) = P(A|B)\,P(B) = P(B|A)\,P(A) \tag{5.7}$$

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 in Figure 5.3 again, note that

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

Using formula (5.7), we can express the joint probabilities $P(A$ and $B)$ by $P(A|B)\,P(B)$. Therefore,

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

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

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n) \tag{5.8}$$

## EXAMPLE 5.15 Using the Multiplication Law of Probability

Texas Holdem 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 $P(\text{ace on first card and ace on second card}) = P(\text{ace on second card} \mid \text{ace on first card}) \times P(\text{ace on first card})$. 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

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

In Example 5.14, 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* $P(A \mid B) = P(A)$.

## EXAMPLE 5.16 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 $P(B_1 \mid M) = 0.397$, $P(B_1)$ was shown to be 0.34 in Example 5.11; 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 (5.7) by substituting $P(A)$ for $P(A \mid B)$:

$$P(A \text{ and } B) = P(A)P(B) = P(B)\,P(A) \tag{5.9}$$

## EXAMPLE 5.17 Using the Multiplication Law for Independent Events

Suppose $A$ is the event that a sum of 6 is first rolled on a pair of dice and $B$ is the event of rolling a sum of 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 $P(A \text{ and } B) = P(A)P(B) = \left(\frac{5}{36}\right)\left(\frac{4}{36}\right) = \frac{20}{1296}$.

### CHECK YOUR UNDERSTANDING

1. Define the terms experiment, outcome, and sample space.
2. Explain the difference between a permutation and a combination.
3. Give an example of each of the three definitions of probability.
4. What are the two key facts that govern probability?
5. What is an event? Explain how to compute $P(A \text{ or } B)$ for two events $A$ and $B$.
6. Explain the concepts of joint, marginal, and conditional probability, and independent events.

## Random Variables and Probability Distributions

Some experiments naturally have numerical outcomes, such as a sum of 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 5.18 Discrete and Continuous Random Variables

The outcomes of the sum 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. Examples of continuous random variables are the weekly change in the DJIA, 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 5.19 Probability Distribution of Dice Rolls

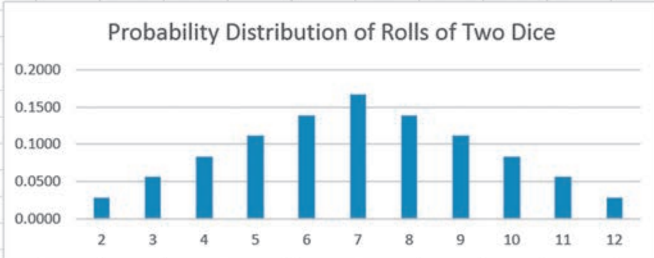
The probabilities of the sum of outcomes for rolling two dice are calculated by counting the number of ways to roll each sum divided by the total number of possible outcomes. These, along with an Excel column chart depicting the probability distribution, are shown from the Excel file *Dice Rolls* in Figure 5.6.

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 5.2) 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.

► **Figure 5.6**

*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 |




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

## EXAMPLE 5.20 A Subjective Probability Distribution

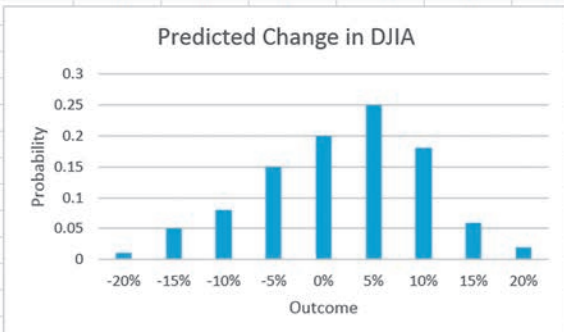
Figure 5.7 shows a hypothetical example of the distribution of one expert's assessment of how the DJIA might change in the next year. This might have been created purely by intuition and expert judgment, but we hope it would be supported by some extensive analysis of past and current data using business 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 5.7**

*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 |



## CHECK YOUR UNDERSTANDING

1. Explain the difference between a discrete and a continuous random variable, and give an example of each.
2. What is a probability distribution?
3. What is an empirical probability distribution?
4. Why is it important to understand the common types of probability distributions?

# 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 $i$th value of the random variable $X$, and $f(x_i)$ is the probability associated with $x_i$.

### EXAMPLE 5.21 Probability Mass Function for Rolling Two Dice

For instance, in Figure 5.6 for the dice example, the values of the random variable $X$, which represents the sum of the rolls of two dice, are $x_1 = 2$, $x_2 = 3$, $x_3 = 4$, $x_4 = 5$, $x_5 = 6$, $x_6 = 7$, $x_7 = 8$, $x_8 = 9$, $x_9 = 10$, $x_{10} = 11$, and $x_{11} = 12$. 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, inclusive, and (2) the sum of all probabilities must add to 1; that is,

$$0 \le f(x_i) \le 1 \quad \text{for all } i \tag{5.10}$$

$$\sum_i f(x_i) = 1 \tag{5.11}$$

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 $P(X \le x)$ and reads as "the probability that the random variable $X$ is less than or equal to $x$."

## EXAMPLE 5.22 Using the Cumulative Distribution Function

The cumulative distribution function for the sum of rolling two dice is shown in Figure 5.8, 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 0.4167. 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 0.4167, then the probability of the complementary event (rolling a 7 or more) is $1 - 0.4167 = 0.5833$. 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, $P(4 \le X \le 8)$, we can find $P(X \le 8)$ and subtract $P(X \le 3)$; that is,

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

*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 $P(X \le 3)$, not $P(X \le 4)$.

## 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:

$$E[X] = \sum_{i=1}^{\infty} x_i f(x_i) \qquad \textbf{(5.12)}$$

Note the similarity to computing the population mean using formula (4.16) 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.

▶ **Figure 5.8**

*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 | |