3.7 The Importance of Scoring First in Soccer

Conditional probabilities can be used to incorporate additional information in a probability calculation, as discussed in previous sections. In this section, we give an example of this to quantify the importance of scoring first in soccer games played in the English Premier League in the 2010-2011 through 2012-2013 seasons. The data analyzed here are available on SoccerSTATS.com.

Based on these data, the probability that the home team wins the game is 0.453. Note that this is an average value for the entire league; we can think of it as applying to a game played between two "randomly chosen" team in the Premier League. We can write this result as

$$P(\text{home team wins}) = 0.453$$

Now, suppose that we include the additional information that the home team scores first. This additional information changes the probability that the home team wins the game. For the Premier League, the conditional probability that the home team wins given that the home team scores first is 0.718. We write this as

$$P(\text{home team wins} \mid \text{home team scores first}) = 0.718$$

Therefore, the fact that the home team scores first has an important effect on the probability that the home team wins the game, changing it from the unconditional probability 0.453 to 0.718. If the visiting team scores first, the probability that the home team wins is 0.178,

$$P(\text{home team wins} \mid \text{visiting team scores first}) = 0.178$$

Again, the fact that the visiting team scores first is important information that greatly affects the probability that the home team wins. There is the possibility that neither team scores first; that is, there are no goals scored; in this case, the game is a draw so that

$$P(\text{home team wins} \mid \text{neither team scores first}) = 0$$

Note that these three conditional probabilities are related. The probability that the home team scores first is 0.534, the probability that the visiting team scores first is 0.390, and the probability that neither team scores first is 0.076. The law of total probability discussed in the previous section can be extended to apply to three conditioning events:

$$\begin{aligned} P(\text{home team wins}) = {} & P(\text{home scores first})P(\text{home team wins} \mid \text{home scores first}) \\ & + P(\text{visitor scores first})P(\text{home team wins} \mid \text{visitor scores first}) \\ & + P(\text{neither scores first})P(\text{home team wins} \mid \text{neither scores first}) \end{aligned}$$

Using the probability values based on Premier League, this expression becomes

$$(0.534)(0.718) + (0.178)(0.390) + (0.076)(0) = 0.453$$

Note that most of the home team's wins come in games in which they have scored first. To express this idea, we write

$$P(\text{home team scores first} \mid \text{home team wins}) = 0.847$$

As discussed in the previous section, this probability is fundamentally different from the probability

$$P(\text{home team wins} \mid \text{home team scores first}) = 0.718$$

The probability 0.847 refers to the fact that in 84.7% of the games in which the home teams wins, the home team scores first. That is, it refers to a proportion of games in which the home team wins. The probability 0.718 refers to the fact that in 71.8% of the games in which the home team scores first, the home team wins the game. That is, it refers to a proportion of games in which the home team scores first.

3.8 The Binomial Distribution

Although the probability distribution of a random variable can be quite general, subject to some simple requirements such as the set of all probabilities summing to 1, in practice, there are few distributions that are particular useful. In this section, we discuss the binomial distribution.

Consider an experiment and let A be an event of interest. Define a random variable X such that $X = 1$ if A occurs and $X = 0$ otherwise. Then,

$$P(X = 1) = P(A)$$

which we can denote by π, where $0 < \pi < 1$. Let $X_1, X_2, \ldots, X_n$ be independent random variables, each with the distribution of X. Then, $X_1, X_2, \ldots, X_n$ is a sequence of ones and zeros. Let

$$S = X_1 + X_2 + \cdots + X_n$$

Note that S is simply the number of times that A occurs in the n experiments. It follows that S is a random variable, and its distribution can be determined using the information provided. S is said to have a *binomial distribution* with parameters n, π. We write

$$S \sim Bin(n, \pi)$$

For instance, suppose $n = 2$. To find the probability that $S = 2$, we need to find all combinations of X_1 and X_2 that yield a sum of 2 and add their probabilities. Because there is only one way to have $S = 2$ (both X_1 and X_2 must be 1), this case is particularly simple:

$$P(S = 2) = P(X_1 = 1, X_2 = 1)$$

because X_1 and X_2 are independent, we have

$$P(X_1 = 1, X_2 = 1) = P(X_1 = 1) \times P(X_2 = 1) = \pi \times \pi = \pi^2$$

On the other hand, finding the probability that $S = 1$ is a little more complicated because there are two ways to have $S = 1$: $X_1 = 1$ and $X_2 = 0$ or $X_1 = 0$ and $X_2 = 1$. Thus,

$$\begin{aligned} P(S = 1) &= P(X_1 = 1, X_2 = 0) + P(X_1 = 0, X_2 = 1) \\ &= P(X_1 = 1)P(X_2 = 0) + P(X_1 = 0)P(X_2 = 1) \\ &= \pi(1 - \pi) + (1 - \pi)\pi \\ &= 2\pi(1 - \pi) \end{aligned}$$

Because

$$P(S = 0) + P(S = 1) + P(S = 2) = 1$$

It is clear that $P(S = 0) = (1 - \pi)^2$. The important assumptions of the binomial distribution are that the results of individual experiments are independent and that the probability that $X = 1$; that is, the event of interest occurs, is the same in each experiment.

The situations in which the binomial distribution applies are quite simple; we identify an event of interest and simply count how often the event occurs. However, it is because of that simplicity that the binomial distribution is so useful. Even when the experiment itself is complicated, we are often interested in relatively simple features of the results.

For example, if our experiment is an NFL season, the detailed results of that experiment would fill this course lecture notes. However, suppose we are interested in whether or not the team with the leading rusher during the regular season wins the Super Bowl. Then, the number of seasons in the past 20 years in which the team with the leading rusher wins the Super Bowl can be modeled as a binomial random variable.

The only quantities governing the binomial distribution are n, the number of experiments and π, the probability of the event of interest occurring in a given experiment. Therefore, all the properties of a binomial random variable S are function of n and π. The expected value and standard deviation of $S \sim Bin(n, \pi)$ are

$$E(S) = n\pi \quad \text{and} \quad \text{SD}(S) = \sqrt{n\pi(1-\pi)} \tag{3.8}$$

For example, if we observe 100 experiments and in each one the probability of A is 0.25, we expect 25 occurrences of A. The form of the standard deviation may seem a little strange but, after a little reflection, it should make sense. The standard deviation is a measure of variation. Suppose π is very close to 0. Then A almost never occurs. Therefore, S is almost always 0; that is, there is very little variation in S. The same argument applies if π is very close to 1, except that A almost always occurs and S is almost always n. That is, when π is close to either 0 or 1, then the standard deviation should be small. We expect a lot of variation whenever $\pi = 1/2$ because A and "not A" are equally likely.

3.9 The Normal Distribution

The second important distribution that we will consider is the *normal distribution*. Unlike the binomial distribution, the normal distribution is a continuous distribution, and if a random variable X has a normal distribution, X can take any value between $-\infty$ and ∞, although extreme values are unlikely.

The normal distribution is governed by two parameters, traditionally denoted by μ and σ. Here, μ represents the mean of the distribution of X, and σ represents the standard deviation; because standard deviations are always positive, $\sigma > 0$. We write

$$X \sim N(\mu, \sigma)$$

The shape of the distribution is given by the well-known bell-shaped curve, which takes its maximum value at μ; σ governs how spread out the curve is. Figure 3.3 shows a few normal distribution, corresponding to different values of μ and σ. These plots illustrate some important properties of the normal distribution. For instance, the distribution is symmetric about its peak, which occurs at the mean of the distribution. When the value of μ changes, the effect on the distribution is a shift; other aspects of the distribution, such as it "bell-shape," don't change. When the value of σ changes, the effect is essentially to change the scale on the x-axis.

Although it is easy to describe the shape of the normal distribution, it is a lite more difficult to determine probabilities associated with a normal distribution. Let X denote a random variable with a normal distribution with mean μ and standard deviation σ. Since X is a continuous random variable, we can't give a table listing the possible values of X together with their probabilities. Instead, we consider the probability that X falls into a certain range of values.

① $\pi = 70\%$, $n = 5$

Let X denote the number of games that the Golden State Warriors win the first five game

(a) $X \sim Bin(5, 0.7)$

(b) $P(X=4) = \binom{5}{4} 0.7^4 (1-0.7)^{5-4}$

$= (5)(0.7)^4 (0.3)^1$

$= 36\%$

(c) $E(X) = 5 \times 0.7 = 3.5$

(d) $Var(X) = 5 \times 0.7 \times (1-0.7)$

$= 5 \times 0.7 \times 0.3$

$= 1.05$

(2) (a) $X \sim Bin(3, 0.46)$

(b) $P(X=2) = \binom{3}{2} 0.46^2 (1-0.46)^{3-2}$

$= (3)(0.46^2)(0.54)^1$

$= 34.3\%$

(c) $E(X) = (3)(0.46) = 1.38$

(d) $SD(X) = \sqrt{3(0.46)(1-0.46)} = 0.863$

In general,

$$X \sim Bin(n, \pi)$$

$$P(X = k) = \underbrace{\binom{n}{k}} \pi^k (1-\pi)^{n-k}$$

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

$$n! = 1 \cdot 2 \cdot 3 \cdot 4 \cdots n$$

$$0! = 1$$

$$5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$$