3.3 Modeling The Results of Sporting Events as Random Variables

The basic rules of probability are concerned with events, which can describe any possible specific outcome that might occur when an experiment is performed. However, in applying analytic methods to sports, we are generally concerned with data; that is, we analyze numbers, not events. Random variables provide the mathematical link between probability theory and data.

A random variable is simply a numerical quantity derived from the outcome of an experiment. Let X denote the number of touch-downs passes thrown by Aaron Rodgers in a Bear-Packers game; X is an example of a random variable. Once the outcome of an experiment is available, that is, once the game is played, the value of X for that experiment is known.

A random variable can be used to define events. For instance, in the example, “Rodgers throws one touchdown pass” is an event; in terms of X, it can be written “$X = 1$.” Therefore, the probability $P(X = 1)$ has the usual interpretation of a long-run frequency. Obviously, there is nothing special about the value 1 in this context, and we might consider $P(X = x)$ for any possible x; note that here X denotes a random variable and x denotes a possible value of X. The set of values $P(X = x)$ for all possible x is called the *probability distribution* of the random variable.

For

~~Fro~~ instance, in the example, X might follow the distribution given in Table 3.1; the values in this table roughly correspond to Rodger’s career regular season statistics for the games he started through the 2012 season. Note that the probability sum to 1, a requirement for a probability distribution.

Table 3.1: Example of a Probability Distribution

x	$P(X = x)$
0	0.10
1	0.25
2	0.25
3	0.25
4	0.10
5	0.05

⊕ = 100%

Events based on random variables, and their probabilities, follow the same rules as other events. Thus, in the example,

$$P(X \le 1) = P(X = 0 \overset{\cup}{\text{or}} X = 1) = \overset{0.10}{P(X = 0)} + \overset{0.25}{P(X = 1)} = 0.10 + 0.25 = 0.35$$

The *distribution function* of a random variable X is the function of x given by $P(X \le x)$; it is often denoted by $F(x)$. Using the probability distribution in Table 3.1, the corresponding function is given in Table 3.2. That is, the distribution function is simply the running totals of the probability distribution.

Table 3.2: Example of a Distribution Function

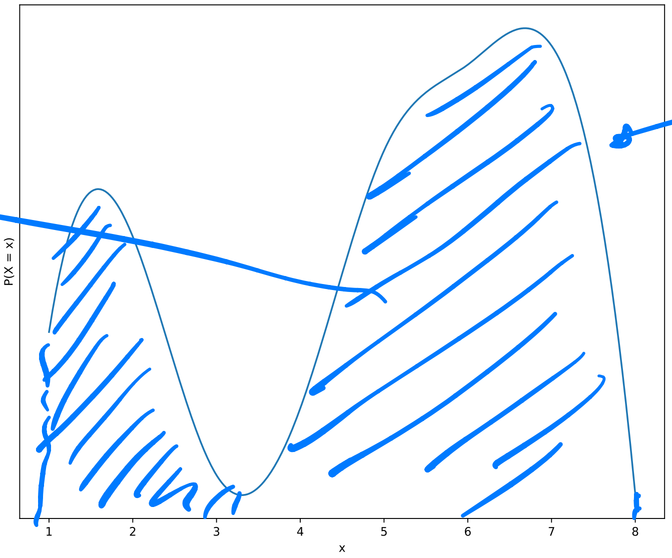
x	$F(x)$
0	0.10
1	0.35
2	0.60
3	0.85
4	0.95
5	1

= 0.25 + 0.10
= 0.25 + 0.25 + 0.10
= 0.25 + 0.25 + 0.25 + 0.10
= 0.10 + 0.25 + 0.25 + 0.25 + 0.10
= 0.05 + 0.10 + 0.25 + 0.25 + 0.25 + 0.10

Random variables, in which the set of possible values may be written as a list, for example: $0, 1, 2, 3, \ldots$ are said to be *discrete*. Hence, the random variable representing Rodgers' touch-downs passes is discrete. The probability distribution and distribution function of a discrete random variable can be given as tables as shown in Tables 3.1 and 3.2. A *continuous* random variable is one that can take any value in a range. Note that the definitions of discrete and continuous random variables are analogous to the definitions of discrete and continuous data discussed in Chapter 2.

It is a little more complicated to describe the probability distribution of a continuous random variable. A useful device for expressing such a distribution is to consider a long-run sequence of experiments. If X is a random variable defined for that experiment, there is a corresponding sequence of random variables $X_1, X_2, \ldots$ such that X_j is based on the j-th experiment. Consider computation of a histogram based on $X_1, X_2, \ldots, X_n$ where n is some very large number. Because n is large, such a histogram could be expressed as a smooth curve such as in Figure 3.1.

area under the curve is equal to 1



probability distribution of a Continuous Random Variable

Figure 3.1: Example of a Hypothetical Histogram

Suppose that the function in Figure 3.1 is standardized so that the total area under the curve is 1. Then the function can be used to express probabilities regarding X. Specifically, for $a < b$, $P(a < X < b)$ can represented by the area under the curve between points a, b as shown in Figure 3.2. In fact, such an approach can be made precise by relating the function in the histogram to the "probability density function" of the random variable and calculating ares using calculus.

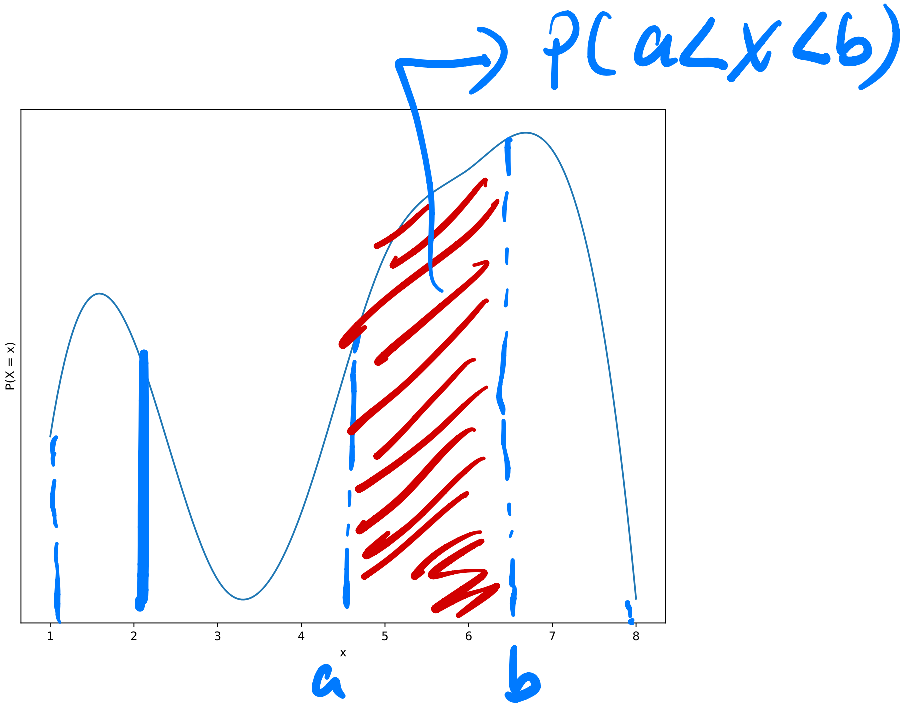


Figure 3.2: $P(a < X < b)$

One consequence of this approach to continuous random variables is that, for any choice of a, $P(X = a)$ is the area under the curve at a. Because, in mathematics, the area of a line is always 0, so $P(X = a) = 0$ for any a.

3.4 Summarizing The Distribution of a Random Variable

The distribution of a random variable can be complex and contains much information. Hence, it is often useful to summarize probability distributions by a few numbers, in the same way that we summarized a dataset in Chapter 2. In fact, any type of summary that can be applied to a dataset can be applied to a random variable by considering random variables $X_1, X_2, \ldots$ obtained from a long sequence of experiments. For instance, the mean of a random variable X can be viewed as the limiting value of the sample mean based on a long sequence of repetitions of the experiment. Therefore, the relationship between the mean of a random variable and a sample mean is the same as the relationship between a probability and a sample frequency. The mean of a random variable is sometimes called its *expected value.*

Consider the example in which X denotes the number of touch-downs passes thrown by Aaron Rodgers; using the distribution in Table 3.1, the mean or expected value of X is 2.15. Therefore, according to this result, in a large number of Bears-Packers games, we expect Rodgers to throw 2.15 touch-downs passes per game.

① $S = \{E_1, E_2, E_3, E_4, E_5\}$

We are given $P(E_1) = 3P(E_2) = 0.3$

Goal : Find $P(E_3)$, $P(E_4)$, $P(E_5)$

$$P(E_1) = \underbrace{3P(E_2) = 0.3}_{}$$

$\Rightarrow P(E_1) = 0.3$ $\Rightarrow \underline{\underline{3}}P(E_2) = 0.3$

$\Rightarrow P(E_2) = \frac{0.3}{3} = 0.1$

We know that

$$P(E_3) = P(E_4) = P(E_5)$$

We also know that

$$\underbrace{P(E_1)}_{0.3} + \underbrace{P(E_2)}_{0.1} + \underbrace{P(E_3)}_{x} + \underbrace{P(E_4)}_{x} + \underbrace{P(E_5)}_{x} = 1$$

$\Rightarrow 0.4 + 3x = 1$

$$\Rightarrow \quad 0.4 + 3X = 1$$

$$\Rightarrow \quad 3X = 0.6$$

$$\Rightarrow \quad X = \frac{0.6}{3} = 0.2$$

$P(E_1) = 0.3$

$P(E_2) = 0.1$

$P(E_3) = P(E_4) = P(E_5) = 0.2$

② ⓐ Define the Sample space

$$S = \{HT, TH, HH, TT\}$$

ⓑ Assign probability to the outcomes

$$P(HT) = \underbrace{P(H)}_{50\%}\underbrace{P(T)}_{50\%} = 25\%$$

$$P(TH) = P(T)P(H) = 25\%$$

$$P(HH) = P(H)P(H) = 25\%$$

$$P(TT) = P(T)P(T) = 25\%$$

ⓒ $A \leftarrow$ one head

$B \leftarrow$ at least one head

$$A = \{HT, TH\}$$

$$B = \{HT, TH, HH\}$$

ⓓ

$$\begin{aligned} P(A) &= P(HT) + P(TH) \\ &= 25\% + 25\% \\ &= 50\% \end{aligned}$$

$$\begin{aligned} P(B) &= P(HT) + P(TH) + P(HH) \\ &= 25\% + 25\% + 25\% \\ &= 75\% \end{aligned}$$

$$\begin{aligned} P(A \cap B) &= P(HT) + P(TH) \\ &= 25\% + 25\% \\ &= 50\% \end{aligned}$$

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 50\% + 75\% - 50\% \\ &= 75\% \end{aligned}$$