Chapter 3

Probability

3.1 Introduction

The goal of analytic methods is to use data to better understand the factors that influence the results of sporting events. However, to do this, we must deal with the fact that all sports have a random element that comes into play; understanding this randomness is crucial to being able to draw the appropriate conclusions from sports data.

Probability theory is the branch of mathematics that deals with random outcomes. This chapter consider the basic properties and rules of probability, focusing on those concepts that are useful in analyzing sports data.

3.2 Applying Probability Rules to Sports

The starting point for probability theory is the concept of an *experiment*. An experiment is simply any process that generates a random outcome. For example, LeBron James shooting a free throw is an example of an experiment.

When analyzing an experiment, we are interested in specific "events." An event is anything that might occur as a result of the experiment. The only requirement is that, based on the result of the experiment, we know whether the event has occurred. For example, Derek Jeter batting against Josh Beckett , "hit," "home run," "strikeout," and "ground out" are all events; however, "Yankees win" would not qualify as an event because, once the bat occurs, we do not know if the Yankees win.

Corresponding to each event of an experiment is its probability. If A is a possible event in an experiment, it is often useful to denote the probability of A by $P(A)$. We can think of $P(\cdot)$ as function that takes different events and returns their probability. For example, in the Jeter-Beckett example, if H denotes "hit," we might write $P(H) = 0.3$ to indicate that the probability that Jeter gets a hit is 0.3.

Although everyone has a good intuitive notion of what a probability is, to use probabilities in a formal analysis we need to have a precise definition. The probability of an event is usually defined as "*long-run relative frequency*." For example, when we say that $P(H) = 0.3$ in the Jeter-Beckett example, we mean that if Jeter faces Beckett in a long sequence of bats, he will get a hit about 30% of the time.

Note that this idea is a hypothetical one. For instance, consider the experiment that Bear play the Packers at home; if we say that the probability that the Bears win is 0.25, we mean that in a long sequence of such games the Bears would win about 25% of the time. However, it is impossible to play a long sequence of games under the same conditions; players would age, some would be injured, coaches would change strategy based on what worked and what did not work, and so on. Thus, the "long-run relative frequency" interpretation is just a way to think about probabilities, not a procedure for determining them.

Probabilities follow some basic rules. Consider an experiment and let A and B be events with probabilities, $P(A)$ and $P(B)$, respectively. It is often useful to refer to results in either A or B occurs, which we write as "A or B." If A and B can't occur simultaneously, then

$$P(A \text{ or } B) = P(A) + P(B)$$

For instance, in the Jeter-Beckett example, let S denote the event that Jeter hits a single and let D denote the event in which Jeter hits a double. If $P(S) = 0.2$ and $P(D) = 0.05$, then the probability that he hits either a single or a double is

$$P(S \text{ or } D) = P(S) + P(D) = 0.2 + 0.05 = 0.25$$

Let H denote the event in which Jeter gets a hit (of any type) and suppose $P(H) = 0.3$. Note that because a single is a hit, S and H can occur simultaneously; hence, we can't calculate $P(S \text{ or } H)$ by summing $P(S)$ and $P(H)$.

A simple, but surprisingly useful, rule applies when we are interested in the probability that an event A *does not* occur; we denote such an event by "*not* A." For example, in the Jeter-

Beckett example, "not S" is the event that Jeter does not hit a single. If the probability that Jeter hits a single is 0.20, then the probability that he does not hit a single must be 0.80. In general,

$$P(\text{not } A) = 1 - P(A)$$

The likelihood of an event is most commonly expressed in terms of its probability, as we have done so far in this section. However, in some cases, it is more convenient to use *odds* rather than probabilities. Consider an event A with probability $P(A)$. The odds of A occurring or, equivalently, the odds in favor of A, are given by the ratio of the probability that A occurs to the probability that A does not occur

$$\text{odds of } A = \frac{P(A)}{1 - P(A)}$$

For example, if A has probability 0.5, then the odds in favor of A are 1 to 1. If A has probability 0.75, then the odds of A are 3 to 1; often, we drop the "to 1" in the statement and say simply that the odds of A are 3. Note that we could also talk about the odds against an event, as is often done in gambling; such odds are given by

$$\frac{1 - P(A)}{P(A)}$$

It is often convenient to use odds against an event when discussing events that have very small probabilities. There are several reasons why odds may be more convenient to use than probabilities. One is that, for very large or very small probabilities, odds are often easier to understand. For instance, if an event has probability 0.00133, in describing the likelihood of this event, it might be more meaningful to say that the odds against the event are about 750 to 1; that is, the event occurs about once in every 750 experiments. Another reason is that, when making a statement about the relative probability of events, odds are often easier to interpret. For instance, the statement that the probability of one event is two times the probability of another can mean very different things if the smaller probability is 0.01 or if it is 0.5; furthermore, such a relationship is impossible if the smaller probability exceeds 0.5. On the other hand, if the odds in favor of an event double, the interpretation tends to be more stable over the range of possible values.

Finally, in the context of sports, odds often better represent the relative difficulty of certain achievements. For instance, consider an NFL quarterback and let A denote the event that he throws an interception on a given pass attempt. Based on 2012 data, for Mark Sanchez, $P(A) = 0.04$, for Sanchez to decrease this by 0.01, he would need to throw 9 interceptions in every 300 attempts. For Tom Brady, the probability of A is 0.013. For Brady to decrease this by 0.01, he would need to be almost perfect, throwing only about 1 interception in every 300 attempts. In terms of odds, changing $P(A)$ from 0.04 to 0.03 is equivalent to changing the odds against an interception from 24 to 49. Changing $P(A)$ from 0.013 to 0.003 is equivalent to changing the odds against an interception from 76 to 333. Therefore, the odds better reflect the fact that, practically speaking, the difference between interception probabilities of 0.04 and 0.03 is larger than the difference between interception probabilities of 0.013 and 0.003.

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.

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

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

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

The *distribution function* of a random variable X is the function of x given by $P(X \leq x)$; 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

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.

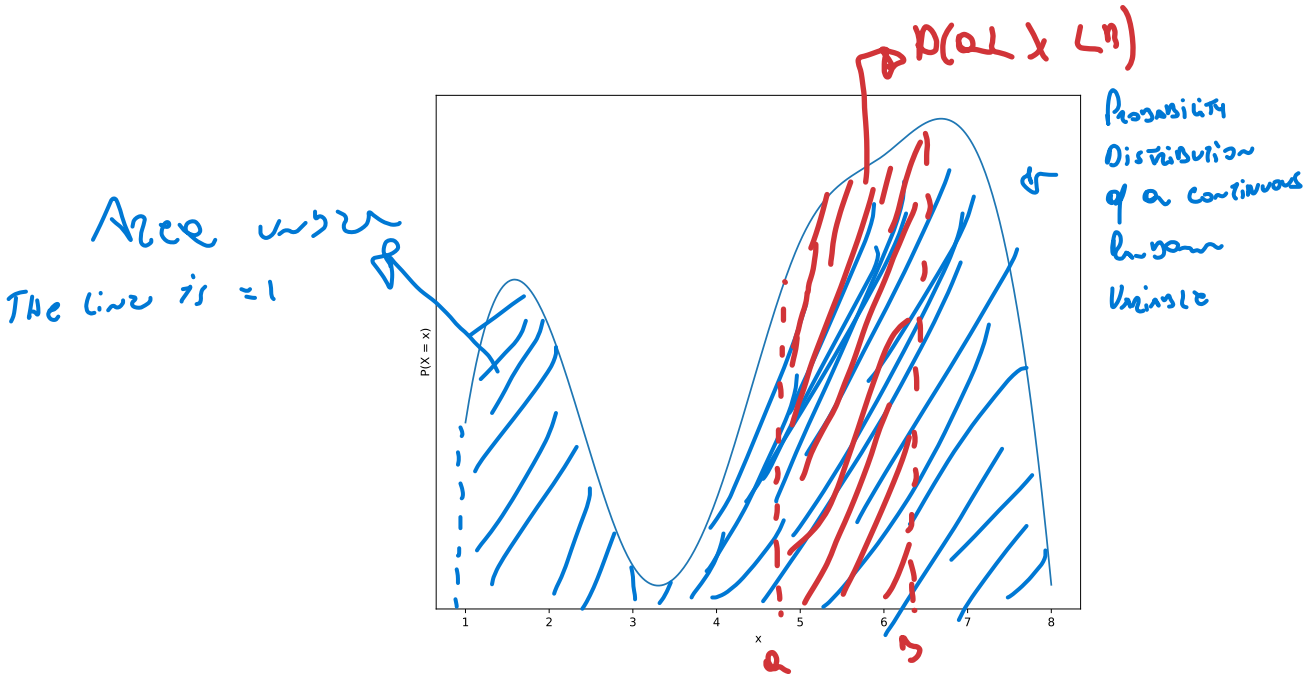


Figure 3.1: Example of a Hypothetical Histogram

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.

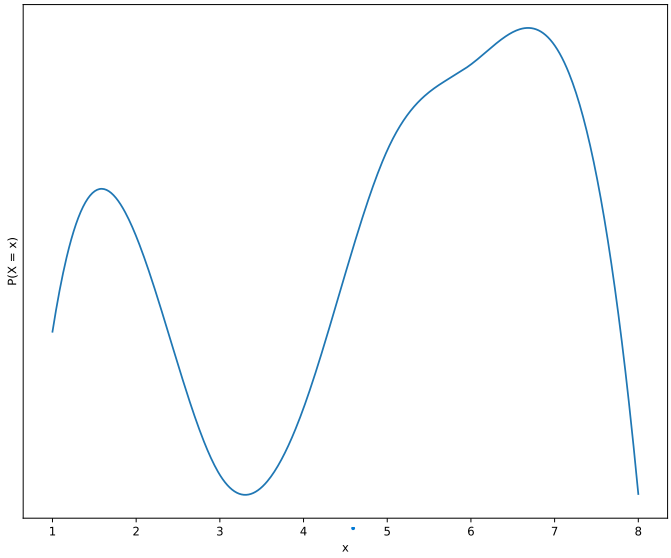


Figure 3.2: $P(a < X < b)$

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.

In the above example, X denotes the number of touch-downs passes thrown by Rodgers, $E(X) = 2.15$. This notation is particularly convenient when describing properties of means. For instance, if X and Y are random variables, then

$$E(X + Y) = E(X) + E(Y) \tag{3.1}$$

That is, the average value of a sum of two random variables is simply the sum of the average values. For example, let Y denote the number of rushing touch-downs scored by Rodgers in a given game and assume that $E(Y) = 0.23$; this corresponds to Rodgers' career average through the 2012 season. Then, Rodgers' average total number of touchdown is $2.15 + 0.23 = 2.38$. More generally, if a, b, and c are constants, then

$$E(aX + bY + c) = aE(X) + bE(Y) + c \tag{3.2}$$

For instance, in the Rodgers example, let $P = 6X + 6Y$ denote the total number of points results from his touch-down passes and rushing touch-downs. Then, M. 18

The same approach used to define the mean of a random variable can be used to define the median, standard deviation, and variance of a random variable. These quantities can all be calculated from a random variable's probability distribution. For example, in the Rodgers' example, it can be shown that X has median 2, standard deviation 1.31, and variance 1.73.

3.5 Conditional Probability

Probabilities are interpreted as long-run relative frequencies in a large sequence of experiments. For example, in the Bears-Packers experiment in which the Bears play the Packers at home, if the event that the Bears win has probability 0.25, this means that in a hypothetical long sequence of games, the Bears will win about 25% of the time. An important part of this interpretation is that the 25% applies to all of the experiments in the sequence. However, in some cases, we might only be interested in those experiments satisfying some further conditions.

Continuing the example, let B be the event that the Bears win so that $P(B) = 0.25$. Now, consider another event, the one in which jay Cutler throws 4 interceptions; denote this event by C. Suppose we are interested in the Bears' probability of winning in those games in which Cutler throws 4 interceptions. We can describe this probability as "the probability that the Bears win *given that* Cutler throws 4 interceptions;" symbolically, we write

$$P(B|C)$$

where the vertical line is read "given that." Probabilities such as $P(B|C)$ are called *conditional probability* because they include additional conditions. Note that, in $P(B|C)$, we are interested in the probability of B; C simply describe the conditions under which the probability is to be calculated. Thus, if, for example, $P(B|C) = 0.05$, then in a long sequence of games *in which Cutler throws 4 interceptions*, the Bears win about 5% of the time. We might write this conditional probability more informally as

$$P(\text{Bears beat the Packers} \mid \text{Cutler throws 4 interceptions}) = 0.05$$

Conditional probabilities are useful because they allow us to incorporate additional assumptions, or additional information, into the probability calculation. Conditional probabilities can be determined from standard, unconditional probabilities. To see how this can be done, consider how we would calculate the conditional in the example if we had access to the results from the long sequence of games.

We are interested in those games in which Cutler throws 4 interceptions, so if we are reading a sequence of game summaries, one per page; we place those in which Cutler threw 4 interceptions in a separate pile, a C pile. We now go through that pile and put those that the Bears won in a second pile. The probability $P(B|C)$ is the ratio of the number of games in the second pile to the number of games in the C pile. Note that the second pile consists of those games that Cutler throws 4 interceptions *and* the Bears win; denote this event by "B and C." Thus, in probability

terms,

$$P(B|C) = \frac{P(B \text{ and } C)}{P(C)}$$

In general, consider an experiment and let A, B be events. Let A *and* B be the event in which both A and B occur. Then,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{3.3}$$

Note that this relationship can be written as

$$P(A \text{ and } B) = P(A|B)P(B) \tag{3.4}$$

so that it gives a formula for finding the probability that two events both occur. That is, the probability that both A, B occur is the probability that B occurs times the probability that A occurs given that B occurs. Note that, because A *and* B, and B *and* A mean the same thing, we also have

$$P(A \text{ and } B) = P(B|A)P(A) \tag{3.5}$$

giving two options for determining $P(A \text{ and } B)$. The conditional probability $P(A|B)$ can be viewed as an "updated" version of the probability, updated to take into account that B occurs. In many cases, this additional information is important in assessing the probability of interest. For instance, in the Bears-Packers example, knowing that Cutler throws 4 interceptions will change the probability that the Bears win. However, in other cases, knowing that B occurs will not change the probability that A occurs. For instance, if K is the event that the Bears kick off to start the game, then it may be reasonable to assume that

$$P(B|K) = P(B)$$

that is, the fact that the Bears kick off to start the game does not change their probability of winning the game. Events A, B for which $P(A|B) = P(A)$ are said to *independent*. For independent events, knowledge about one of them does not change our probability of the other. By rearranging the formula for $P(A|B)$, it can be shown that A, B are independent if and only if

$$P(A \text{ and } B) = P(A)P(B) \tag{3.6}$$

That is, for independent events A and B, the probability that both occur is simply the product of the individual probabilities. Consideration of conditional probabilities shows that it is important to be aware of the conditions under which a probability is calculated and to pay close attention

to the exact probability considered. In particular, care is needed when interpreting evidence presented in the form of conditional probabilities. This issue is illustrated in the example that follows.

In 2011, Andre McCutchen had 62 extra-base hits; only 12.9% of these were with 2 or more runners on base (data of this type is available in Baseball-Reference.com) Does this suggest that McCutchen does not hit as well with runners on base? Note that, in MLB in general for 2011, 13.9% of extra-base hits occurred with two or more runners on base.

In analyzing the meaning of results like these, it often useful to express them using probability notation. The experiment here is a McCutchen at bat, and the events of interest are "had an extra-base hit" and "2 or more runners on base." Note that even though 2 or more runners on base is not a direct consequence of McCutchen's at bat, when the at bat occurs, we know the number of runners on base, we we can view it as an event.

According to 2011 data, 12.9% of the time McCutchen has an extra-base hit, there are 2 or more runners on base. In probability notation,

$$P(\text{2 or more runners are on base} \mid \text{McCutchen has an extra-base hit}) = 0.129$$

It follows that

$$P(\text{less than 2 runners are on base} \mid \text{McCutchen has an extra-base hit}) = 0.871$$

because 87.1% of the time he has an extra-base hit there are 0 and 1 runners on base. Therefore, 12.9% refers to the tendency of there being 2 or more runners on base when McCutchen has an extra-base hit. Therefore, when McCutchen has an extra-base hit, it is relatively unlikely that there are at least 2 base runners. Note, however, that these values do not directly assess McCutchen's tendency to have an extra-base hit with 2 or more base runners. To do this, we should look at his extra-base hit probability in the two situations. That is, we should compare

$$P(\text{McCutchen has an extra-base hit} \mid \text{2 or more runners on base})$$

and

$$P(\text{McCutchen has an extra-base hit} \mid \text{less than 2 runners on base})$$

In 2011, McCutchen had only 66 at bats with 2 or more runners on base, and in 8 of those at bats he had an extra-base hit. Therefore,

$$P(\text{McCutchen has an extra-base hit} \mid \text{2 or more runners on base}) = \frac{8}{66} = 0.121$$

that is, if he comes to bat with 2 or more runners on base, there is a 12.1% chance he will have an extra-base hit (based on 2011 data). He had 506 at bats with either the bases empty or 1 base runner, and he had 54 extra-base hits. Therefore, based on this data

$$P(\text{McCutchen has an extra-base hit} \mid \text{less than 2 runners on base}) = \frac{54}{506} = 0.107$$

that is, if he comes to bat with fewer than 2 base runners, there is about 10.7% chance that he will have an extra-base hit. It follows that in 2011 McCutchen was actually more likely to have an extra-base hit with 2 or more base runners.

The lesson here is that, in making comparisons of this type, it is important to distinguish between the event of interest (having an extra-base hit) and the event defining the relevant situation (2 or more runners on base) and calculate the probabilities accordingly.

3.6 The Law of Total Probability

There is a simple formula relating unconditional and conditional probabilities. Consider the 2013 St. Louis Cardinals. They won 97 games, for a winning "percentage" of 0.599. However, like most MLB teams, they had a higher winning percentage in home games than in road games. At home, they won 54 of 81 games, for winning percentage of 0.667, while on the road, they won 43 games, for a winning percentage of 0.531.

These results can be expressed in probability notation. Let W be the event that St. Louis wins and let H denote the event that the game is a home game. Then

$$P(W) = 0.599, \quad P(W|H) = 0.667 \quad P(W| \text{ not } H) = 0.531$$

Because the cardinals play the same number of home games as away games, their overall winning percentage is simply the average of their home and away winning percentage:

$$\frac{0.667 + 0.531}{2} = 0.599$$

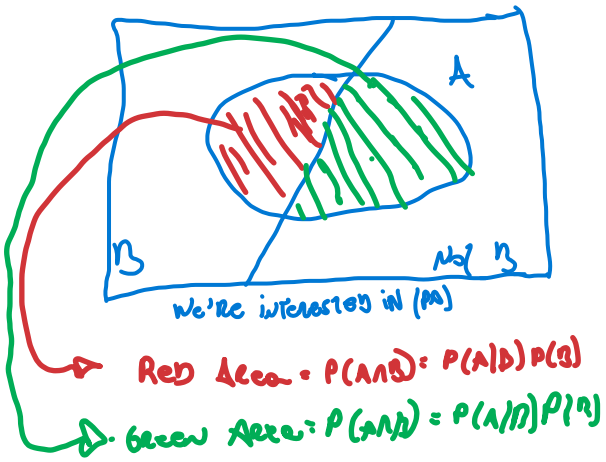
or in probability notation

$$P(W) = P(H)P(W|H) + P(\text{ not } H)P(W| \text{ not } H)$$

because $P(H)$ and $P(\text{ not } H$ are both 0.5. This result, relating unconditional probabilities, is known as the *law of total probability*. Consider an experiment and let A, B denote events. Then, the law of total probability states that

$$P(A) = P(B)P(A|B) + P(\text{ not } B)P(A|\text{ not } B) \tag{3.7}$$

That is, the unconditional probability of an event A can be expressed in terms of a weighted average of the conditional probabilities of A given B and given "not B," where the weights depend on the probability of B.



In the winning percentage example, each MLB team plays 81 home games and 81 away games, so that each team's overall winning percentage is the average of its home and away winning percentages. That is, the relationship between the unconditional probability of the event W has a simple relationship to the conditional probabilities of W given H and W given "not H." Because all teams play 81 home games and 81 away games, the relationship between home and away winning percentages and the team's overall winning percentage is the same for each team. In other cases, the probabilities of the conditioning event may vary for different players or teams, making comparisons more difficult.

Consider the following example. In 2009, Josh Beckett and Johan Santana both had solid years, with similar statistics. In particular, both pitchers had a batting average against (BAA) of 0.244, with Santana holding a slight edge with a value of 0.2438 compared to Beckett's 0.2441. Furthermore, both were much stronger against right-handed batters: Beckett had a BAA of 0.226 against right-handed batters and a BAA of 0.258 against left-handed batters, while Santana had a BAA of 0.235 against right-handed batters and a BAA of 0.267 against left-handed batters.

Therefore, Beckett's BAA was 9 points lower than Santana's against right-handed batters and against left-handed batters. Yet, their overall BAAs was virtually the same, with Santana's slightly lower. This surprising result, called *Simpson's paradox* in statistics, can be explained by the fact that Beckett and Santana faced different proportions of right and left handed batters. Of the 811 at bats against Beckett, only 43.2% were from the right side; of the 640 at bats against Santana, 71.9% were by right-handed batters.

The relationship between the pitchers' overall BAA and their side-specific BAAs follows from the law of total probability. Let H denote the event that, in a given at bat, a pitcher allows a hit. Let R denote the event that the batter is right-handed and let L denote the event that the batter is left-handed; note that L is "not R." Then, the law of total probability states that

$$P(H) = P(R)P(H|R) + P(L)P(H|L)$$

Here $P(H|R)$ is a pitcher's BAA versus right-handed batters, and $P(H|L)$ is his BAA versus left-handed batters. Using Beckett's statistics. this relationship becomes

$$(0.432)(0.226) + (0.568)(0.258) = 0.244$$

while for Santana,

$$(0.719)(0.235) + (0.281)(0.267) = 0.244$$

Therefore, while Beckett was better than Santana versus both right-handed and left-handed batters, Santana faced more right-handed batters than did Beckett, lowering his overall BAA. It follows that the conditional probabilities, in this case their BAA values versus right and left-handed batters, give different information about the relative performance of Beckett and Santana than do the unconditional probabilities, their overall BAA values.

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 pervious 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.

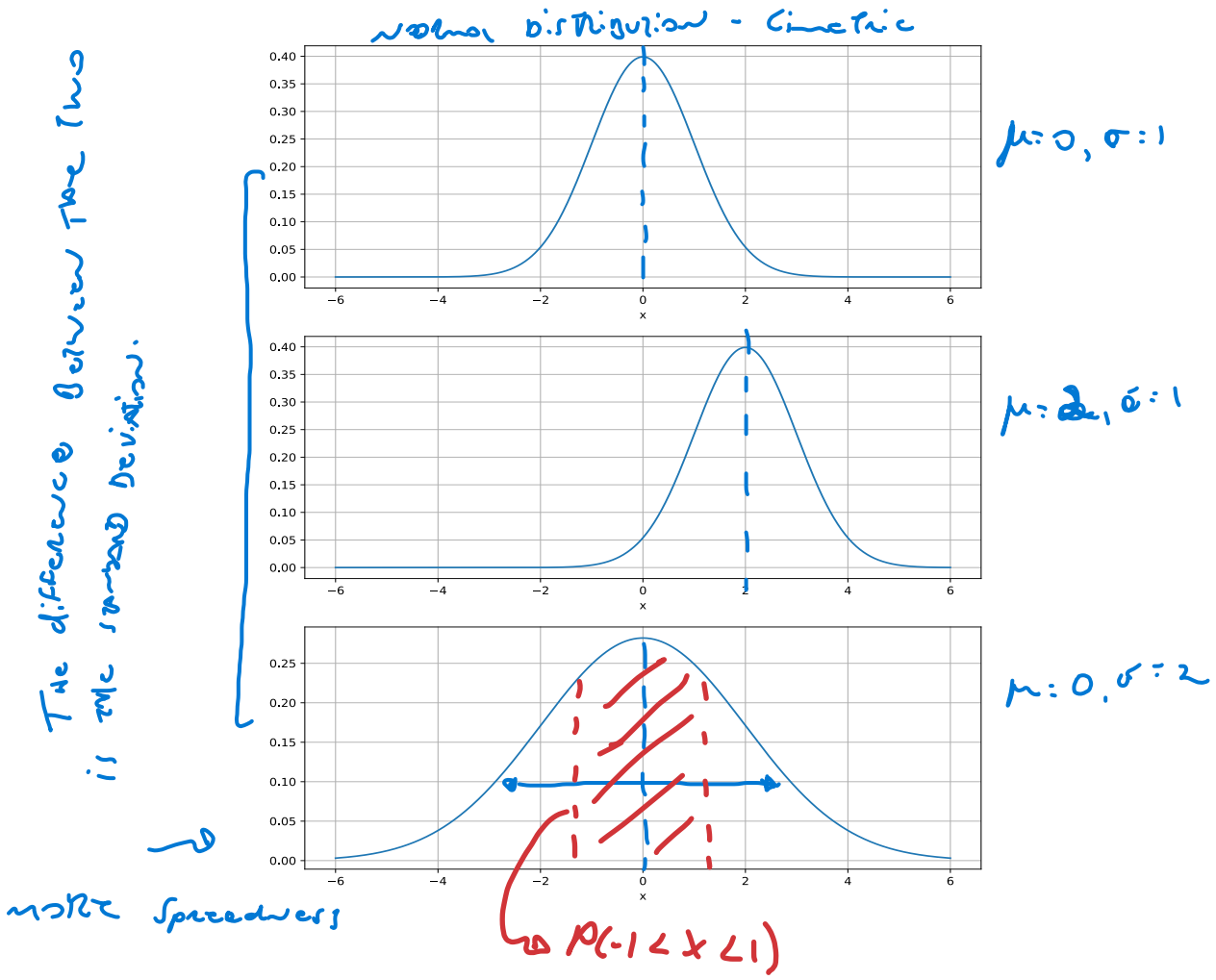


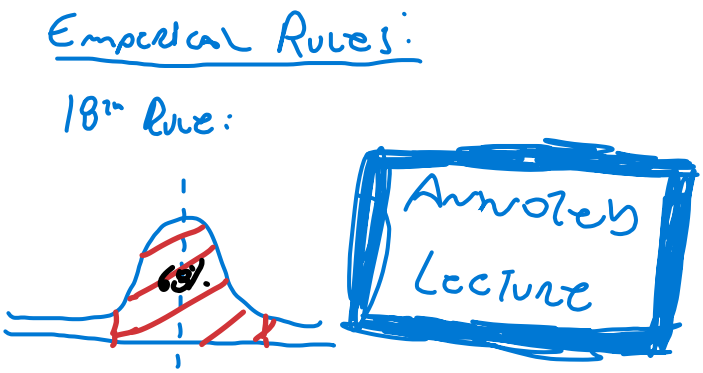
Figure 3.3: Example of normal distributions

To describe such probabilities, it is useful to relate X to a *standard normal distribution*. A standard normal distribution is the one in which $\mu = 0$ and $\sigma = 1$. Let X denote a random variable with mean μ and standard deviation σ. We can convert X to a random variable with a standard normal distribution by computing:

$$Z = \frac{X - \mu}{\sigma} \tag{3.9}$$

Then Z has a standard normal distribution. The standard normal distribution is a convenient reference distribution that can be used to understand variation in many different contexts. Let Z have a standard normal distribution and consider $P(-a < Z < a)$ as a function of a.

Standar normal



Note that $P(-a < Z < a)$ rapidly approaches 1 as a increases. This is an important property of the normal distribution; with high probability, normal random variables tend to be close to their mean value.

Now, consider a normal random variable X that has mean μ and standard deviation σ. To find $P(-b < X < b)$ for some value b, we convert this probability into a probability concerning a standard normal random variable Z. Because $\frac{X-\mu}{\sigma}$ has a standard normal distribution

$$\begin{aligned} P(-b < X < b) &= P\left(-\frac{(b-\mu)}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) \\ &= P\left(-\frac{(b-\mu)}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \end{aligned}$$

Note that $\frac{(b-\mu)}{\sigma}$ is called the Z-score of b. The Z-score of b is the number of standard deviations it is above or below the mean of the distribution. The fact that $P(-1 < Z < 1) = 0.683$ can be

interpreted as saying that the probability that a normal random variable falls within one standard deviation of its mean is 0.683. Therefore, in the statement "with high probability, normal random variables tend to be close to their mean value," *close* is interpreted as being relative to the standard deviation of the random variable.

3.10 Using Z-scores to Compare Top NFL Season Receiving Performance

The Z-score of a measurement Y is defined as

$$\frac{Y - \mu}{\sigma}$$

where μ is the mean of Y values for some distribution, and σ is the corresponding standard deviation. In the previous section, Z-scores were used as a way to understand, and calculate, probabilities regarding random variables with normal distributions. In this section, Z-scores are used to compare and standardize measurements.

The Z-score gives the number of standard deviations a measurement is above or below the mean. For example, if a measurement has a Z-score of 2, then that measurement is 2 standard deviations greater than the average value. Therefore, the Z-score takes into account both the average value of the measurement and the variability of the measurement, as measured by the standard deviation; Z-score give a simple way to compare a statistic for a particular player or team to the values obtained by other players or teams.

In 2012, Calvin Johnson had 1964 receiving yards, the highest yearly total up to that time. It is natural to ask how Johnson's 2012 season compares to other great years for receivers. In Table 3.3 shows the receiving yard totals for Johnson and 5 other receivers; these were chosen to represent a wide range of eras and are not necessarily the 5 best seasons in terms of receiving yards.

Steps for a fair comparison:

Table 3.3: Top receiving yard performances in different eras

Player	Year	Receiving Yards
Calvin Johnson	2012	1964
Marvin Harrison	2002	1722
Jerry Rice	1995	1848
John Jefferson	1980	1340
Otis Taylor	1971	1110
Raymond Berry	1960	1289

- for each player
- championship avg and SD of the year
- then, find the z-score
- whoever has the bigger z-score is the best

Direct comparison of the six seasons totals in Table 3.3 may be misleading because of the way the role of the passing game in the NFL has changed over the years. One way to account for these

differences is to compare each receiver to the other receivers that played that season.

Using the Z-score approach, we convert each player's performance to a Z-score and then compare the Z-scores of the different players; the player with the highest Z-score has the best performance relative to his peers. Let Y_0 be the receiving yards for a player in a given year. For the mean and standard deviation of Y_0, we must use the sample-based values. Therefore, let $\overline{Y}_0$ denote the average receiving yards for some group of players and let S_0 denote the standard deviation of receiving yards for those players. Then, the Z-score of Y_0 is given by

$$\frac{Y_0 - \overline{Y}_0}{S_0}$$

To implement this approach, we need to choose players to use to calculate $\overline{Y}_0$ and S_0. One possibility is to use the set of all players catching at least one pass in the given year. Table 3.4 shows the average and standard deviation of receiving yards for those players for each year represented in Table 3.3. Table 3.4 also shows the Z-scores of the receiving yards for those players in Table 3.3.

Table 3.4: Mean and standard deviation of receiving yards for all players with at least one reception for the years represented in Table 3.3

Year	Average	SD	Z-score
2012	269.2	329.6	5.142
2002	291.2	325.8	4.392
1995	288.4	342.8	3.600
1980	280.5	278.4	3.806
1971	224.3	223.8	3.958
1960	234.0	263.9	4.032

According to the results of Table 3.4, Calvin Johnson's performance in 2012, which corresponds to a Z-score of about 5.1, is the most impressive, with this receiving yards over 5 standard deviations higher than the average receiving yards for players with at least 1 reception.