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.

For example, if I roll a fair die

potential outcomes

1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

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

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

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

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

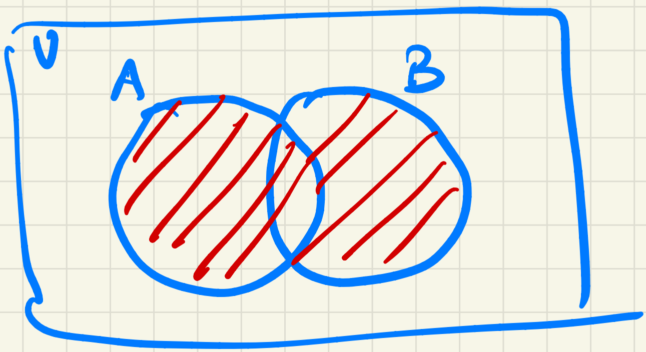
Let H denote the event in which Jeter gets a hit (of any type) and suppose $P(H) = 0.3$. Note that because a single is a hit, S and H can occur simultaneously; hence, we can't calculate $P(S \text{ or } H)$ by summing $P(S)$ and $P(H)$.

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-

In general: Let A & B be events, then

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

Venn - diagram



For any event A, we have

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

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

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,

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

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

$\frac{0.75}{1-0.75} = \frac{0.75}{0.25} = \frac{3}{1}$

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

$\frac{0.5}{1-0.5} = \frac{0.5}{0.5} = \frac{1}{1}$

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

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

not happening

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.