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(A)+ P(not A) =1

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

It follows that

+ 100

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

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,

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

We Know

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

In conditional probability, we have

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

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:

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

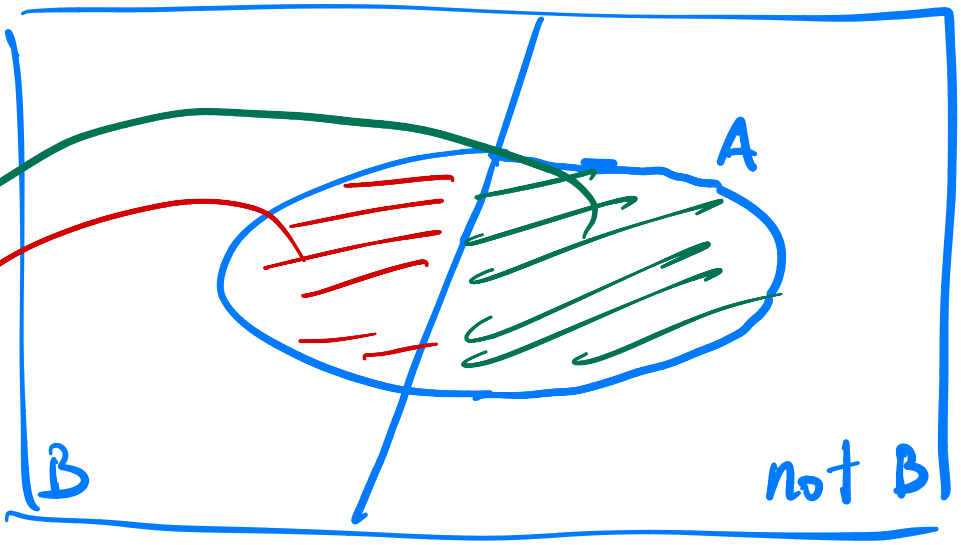
or in probability notation

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

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

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

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.



We are interested in $P(A)$.

Red Area $= P(A \cap B) = P(A|B)P(B)$

green Area $= P(A \cap \text{not } B) = P(A|\text{not } B)\, P(\text{not } B)$

$P(A) =$ Red area + green are

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

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

$$\boxed{P(H) = P(R)P(H|R) + P(L)P(H|L)}$$

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

$$\overset{R}{(0.432)}(0.226) + \overset{L}{(0.568)}(0.258) = 0.244$$

while for Santana,

$$\overset{R}{(0.719)}(0.235) + \overset{L}{(0.281)}(0.267) = 0.244$$

=

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.