1.4.3. Probability and Bayes' Theorem.

We should all be familiar with the fact that the probability of an event can range from zero (impossible) to one (certain) such that a probability near zero means highly unlikely and a probability near one means almost certain. Fifty-fifty means a probability of one half, which is the probability of tossing a head on a fair coin. In general, if there are N equally likely alternatives that are mutually exclusive (meaning they cannot both happen, *e. g.*, a coin cannot land on both faces at the same time) then the probability of each alternative is $1/N$. This result arises because if all possible and distinct outcomes are accounted for, then the probabilities of each distinct outcome must sum to one. This is another way of saying that one of the distinct alternatives is certain to happen.

Rule 1a: *If A and B are distinct, mutually exclusive alternatives then the probability of either A or B is equal to the sum of the probability of A and the probability of B.*

More generally, when A and B may not be mutually exclusive:

Rule 1b: *The probability of either A or B occurring is equal to the sum of the probability of A and the probability of B, minus the probability of them both happening.*

In mathematical form this statement is written as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \qquad \text{Eq. 1.1}$$

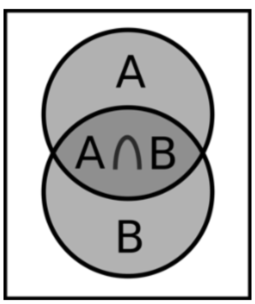

Figure 1.10. In the Venn diagram, left, the entire top circle represents $P(A)$, the entire bottom circle represents $P(B)$, and the overlap is $P(A \cap B)$. The total shaded region is $P(A \cup B)$ whose area is the sum of the areas of each circle minus the area of intersection, which is otherwise counted twice.

The Venn diagram in Figure 1.20 demonstrates this point—the intersection, $P(A \cap B)$ represents the probability of both A and B occurring (the intersection is within both circles). Since the sum of the areas of the two circles would include the intersection twice, the total shaded area is the sum of the two circles minus the area of the intersection.

In examples such as coin tosses and dice rolls, where only one outcome is possible, then $P(A \cap B) = 0$, where A and B refer to any of the distinct results.

For example, if we want to know the probability of rolling a 5 or higher with a standard, 6-faced dice, we should add together the probability of rolling a 5 and the probability of rolling a 6. That is, if n is the number rolled, then

$$P(n \geq 5) = P(n = 5) + P(n = 6)$$
$$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

Our next example will lead us to Rule 2. Suppose we want to know the likelihood of tossing heads three times in a row. A key point in allowing us to solve the problem is the notion of independence—the probability of tossing heads on any one toss does not depend on the results of the other tosses. In this case, since each toss has two equally likely outcomes, there are $2 \times 2 \times 2 = 8$ possible outcomes: HHH, HHT, HTH, HTT, THH, THT, TTH, TTT (where H means heads tossed and T means tails tossed—see Figure 1.11). So, the probability of three heads in a row is $1/8 = 1/2 \times 1/2 \times 1/2$. Rule 2 expresses this result more generally.

Rule 2: *The probability of two or more independent events all occurring is the product of the probabilities of each event occurring individually.*

Mathematically, independence means the probability of one event is unchanged by the occurrence of the other event. More precisely, two events, A and B, are independent if

$$P(A|B) = P(A)$$

where $P(A|B)$ means the probability of A if B occurs, which is said as "the probability of A given B".

Box 1.19. Independent events: Two events are independent of each other if the outcome of one event does not alter the probabilities of any outcome of the other event.

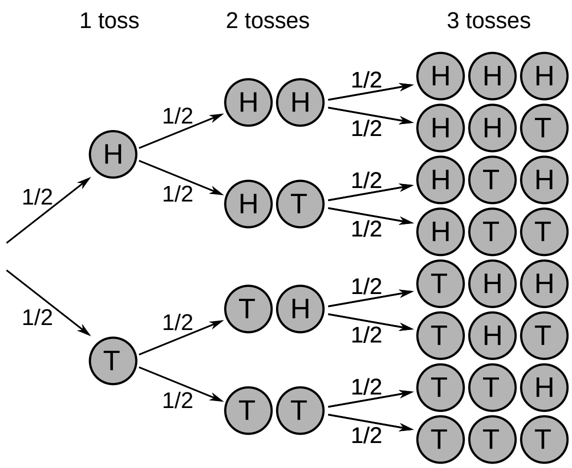

Figure 1.11. Probability of multiple independent events. If a coin is tossed multiple times, the number of possible outcomes for the sequence of tosses doubles following each toss, so there are 8 outcomes after 3 tosses. The probability of Heads or Tails is ½ for every toss, independent of the number of prior Heads or Tails. Each of the final sequences (HHT, HHT, etc.) is equally probable, with a probability of $\frac{1}{2^N}$, where N is the number of coin tosses.

Note on independent events

It is not always clear when events are independent, and intuitively we can make mistakes by forgetting independence. For example, if you had just tossed tails three times in a row you might think that the next toss is more likely to be heads, simply because tossing four tails in a row is very unlikely. However, it is not possible for previous coin tosses to affect a later one, so the probability of heads is still ½. The sequence of results TTTH is exactly as likely as the sequence TTTT. The reason that getting one head is more likely than none when a coin is tossed four times is because there are four ways of getting a single head (HTTT, THTT, TTHT & TTTH). Once we know the first three tosses are all tails, three of the four ways of getting a single heads are no longer available and

$$P(H|TTT) = P(H) = 1/2.$$

Also,

$$P(TTTH) = P(TTT)P(H) = (1/8)(1/2) = 1/16 = P(TTTT).$$

In other cases, we might think that two events are independent and proceed as if they are, in which case our calculations would be incorrect if the events were not independent, meaning they were correlated. For example, the probability of a hot and sunny day is not simply the product of the two separate probabilities, *i. e.,*

$$P(Hot, Sunny) \neq P(Hot) \times P(Sunny),$$

as a day is more likely to be hot if it is sunny, which means the two are positively correlated.

Box 1.20. Positively correlated: Events are positively correlated if the occurrence of one event means the other event is more likely than otherwise. Mathematically, two events, A and B, are positively correlated if $P(A|B) > P(A)$, which also implies $P(B|A) > P(B)$ and $P(A \cap B) > P(A)P(B)$.

> <u>Rule 3:</u> *The probability of any two events occurring is the product of the probability of one event occurring and the probability of the other once we know the first has occurred, or mathematically* $P(A \cap B) = P(A)\ P(B|A)$.

This equation is stated as "the probability of both A and B (or A intersection B) is equal to the probability of A multiplied by the probability of B given A".

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

Probability of A intersection B (Probability of both A and B occurring)

Probability of A occurring

Probability of B given A (Probability of B occurring once we know that A has occurred)

Figure 1.12. Annotation of Rule 3.

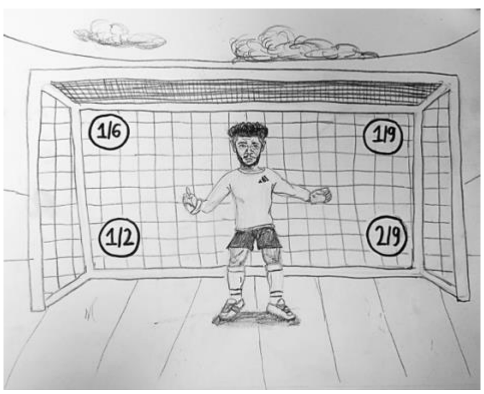

Figure 1.13. Probabilities for each of 4 choices of target by a soccer player taking a penalty. In this example, the probability that the target is bottom-left is $P(B \cap L) = 1/2$. This can be calculated as the probability the player aims left, $P(L) = 1/2 + 1/6 = 2/3$, multiplied by the probability the player shoots low when aiming left, $P(B|L) = \frac{P(B \cap L)}{P(L)} = \frac{1/2}{2/3} = 3/4$. Illustration by Jaleel Draws of Pencil on Paper (POP).

For example, suppose a goalkeeper in a soccer game (Figure 1.13) is facing a penalty kicker who shoots left 2/3 the time and right 1/3 of the time. When the kicker shoots left, the ball goes to the top-left corner 1/4 of the time and the bottom-left corner 3/4 of the time. When the kicker shoots right, the ball goes to the top-right corner 1/3 of the time and the bottom-right corner 2/3 of the time. The goalkeeper has to choose where to dive and hope to be correct in order to save the penalty. The goalkeeper dives to the bottom-left. What is the probability the choice is correct?

> Answer: From Rule 3, probability of a bottom-left is probability the kicker aims leftward, multiplied by the probability the ball goes to the bottom-left when it is aimed leftward. That is, $P(B \cap L) = P(L)P(B|L) = \frac{2}{3} \times \frac{3}{4} = \frac{1}{2}$, so the answer is ½.

You should convince yourself that all 4 probabilities shown in Figure 1.13 can be generated in a similar manner.

Rule 3 can be stated as: The probability of both A and B occurring is the probability of A **and** the probability of B occurring if A occurs. This can be rewritten by symmetry as $P(A \cap B) = P(B)\ P(A|B)$. The second way of writing the equation states the probability of both occurring is the probability of B **and** the probability of A occurring if B occurs. Both of these must be true and both methods of calculating the result must lead to the same answer.

For example, in Figure 1.13, the probability of the penalty-taker aiming low is (1/2 + 2/9) = 13/18. The probability of the penalty-taker aiming left is we know the aim is low, is (1/2)/(13/18) = 9/13. We can combine these to say the probability the shot is bottom-left is the probability the shot is low multiplied by the probability the shot is left when it is low.
That is, $P(B \cap L) = P(L)P(B|L) = \frac{13}{18} \times \frac{9}{13} = \frac{1}{2}$.

<u>*Bayes' Theorem*</u>
We have just seen that Rule 3 can be written in two ways, both as $P(A \cap B) = P(A)\ P(B|A)$ and, by symmetry, $P(A \cap B) = P(B)\ P(A|B)$. Bayes' Theorem simply identifies these two alternative equations as $P(A)\ P(B|A) = P(B)\ P(A|B)$. This is usually rearranged to give:

<u>*Rule 4*</u> *(Bayes' Theorem)*

$$P(B|A) = \frac{P(B)\ P(A|B)}{P(A)}. \qquad \text{Eq. 1.2}$$

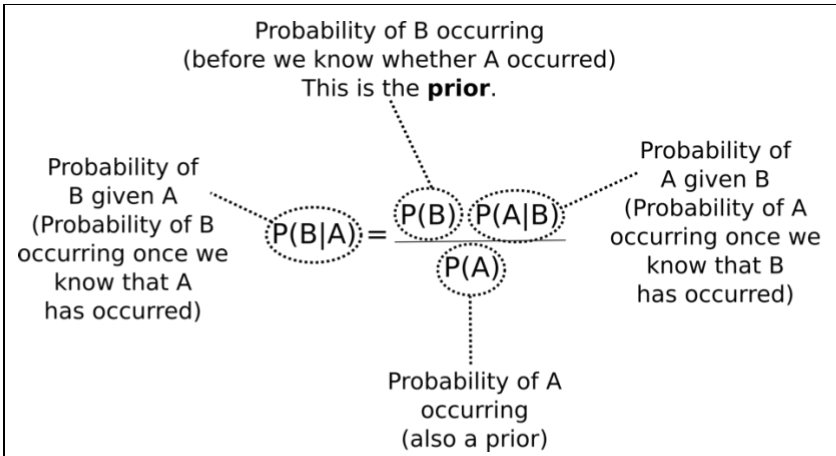

Figure 1.14. Annotation of Bayes' Theorem, Rule 4.

The formula is typically used once we know that Event A has occurred. The terms P(B) and P(A) are priors, as these are respectively the probabilities of Events B and A occurring, before we know whether A occurred.

If A and B are independent then knowing that Event A has occurred has no impact on the probability of Event B occurring, so $P(B|A) = P(B)$. This shows that *Rule 2* is a special case of *Rule 3*, as $P(A \cap B) = P(A)\ P(B|A) = P(A)P(B)$ if A and B are independent.

The test for independence need only be carried out in one direction: if the probability of Event A is not affected by Event B then the probability of Event B is not affected by Event A. This can be seen from Bayes' Theorem (Eq. 1.32, Figure 1.14), as if $P(A|B) = P(A)$ then $P(B|A) = \ P(B)$.

<u>Examples of Bayes' Theorem</u>

(i) Biased coin.

Suppose you have five coins, one of which is biased and produces Heads with a probability of 3/4, while the others are fair so produce Heads with a probability of 1/2 (see Figure 1.15). Otherwise the coins are identical. You choose one of the coins at random, flip it three times and get Heads each time. What is the probability it is the biased coin?

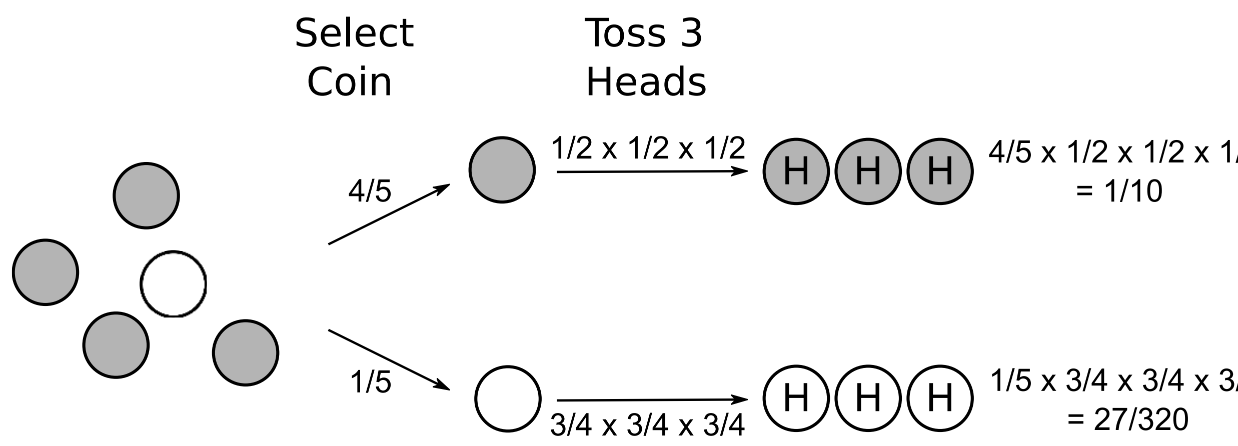

Figure 1.15. Flipping a biased or unbiased coin (Example (i)). The probability of selecting a biased coin, $P(B)$, is 1/5, and the probability of selecting an unbiased coin, $P(\tilde{B})$, is 4/5. The probability of tossing 3 successive Heads when the biased coin is selected, $P(A|B) = \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{64}$, while the probability of tossing 3 successive Heads when the unbiased coin is selected, $P(A|\tilde{B}) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$. Combining these yields the probability of selecting the biased coin and tossing 3 Heads as $P(A \cap B) = P(B)P(A|B) = \frac{1}{5} \times \frac{27}{64} = \frac{27}{320}$, while the probability of selecting the unbiased coin and tossing 3 successive Heads is $P(A \cap \tilde{B}) = P(\tilde{B})P(A|\tilde{B}) = \frac{4}{5} \times \frac{1}{8} = \frac{1}{10}$.

For this problem, we let Event B be “Biased coin is randomly selected” and Event A be “Three Heads in a row” are tossed. We use the nomenclature $\tilde{B}$ to represent the Event “Not B” meaning a fair coin was randomly selected.

In this case, we know or can calculate the following from the question, combined with the above four rules:

$P(B) = 1/5$ (1 out of 5 coins is biased, so probability it is chosen is 1/5).

$P(A|B) = \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{64}$ (if the coin is biased, this is the probability of tossing three Heads in a row from *Rule 2*).

Now $P(A)$ is the probability of tossing three heads in a row, irrespective of what coin is used. This is the sum of the probability of choosing the biased coin and then tossing three heads **plus** the probability of choosing a fair coin and then tossing three heads. Using *Rule 3* this can be written as:

$$P(A) = P(A|B)P(B) + P(A|\tilde{B})P(\tilde{B}) = \left(\frac{1}{5}\right)\left(\frac{27}{64}\right) + \left(\frac{4}{5}\right)\left(\frac{1}{8}\right) = \frac{59}{320}.$$

Finally, we can use Bayes Theorem (*Rule* 4) to evaluate the probability the coin chosen was biased given you tossed three Heads in a row, $P(B|A)$, as

$$P(B|A) = \frac{P(B)\,P(A|B)}{P(A)} = \frac{\left(\frac{1}{5}\right)\left(\frac{27}{64}\right)}{\frac{59}{320}} = \frac{27}{59} = 0.458.$$

You should verify you get the same answer if you use an alternative form (see Figure 1.15):

$$P(B|A) = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap \tilde{B})}.$$

The answer is less than 1/2, which tells us that even though the biased coin is much more likely to produce three Heads in a row than a fair coin, the coin that produced three Heads in a row is more likely to be a fair coin simply because there are more fair coins to choose from.

In this example, the **prior** probability of the biased coin was 1/5. The experimental result of tossing three Heads produced a **posterior** probability of 0.458, which is higher than the prior. If we continued with a new independent test of the coin, we could begin the new experiment assuming the prior probability is 0.458. That is, in a sequence of experiments the posterior probability following one experiment can serve as the prior probability of the next one.

There is evidence that within our brain the experiences of a lifetime act as a prior for interpretation of any new experience. When a stimulus induces a pattern of activity among neurons, connections change between the neurons such that the particular activity-pattern is more likely to be induced by future stimuli. Thus, our brains may have evolved such that we naturally use Bayes' Theorem when responding to the outside world (*cf.* Tutorial 8.4).

(ii) Medical condition.

When doctors examine a patient, they must use Bayes' Theorem, either explicitly or implicitly, to determine the probability of a particular illness given the symptoms. From prior cases, the doctor can know the probability of any symptom when a patient has a certain illness. Similarly, the doctor can know how common an illness is and perhaps form an appropriate prior—*i. e.*, the probability the patient has an illness before carrying out any examination or test—by taking into account the patient's age, gender and family history. Moreover, if the patient arrived in the middle of an outbreak of one disease then the prior probability of that disease would increase. Finally, although precise probabilities are not known in the way they might be in games with coins or cards, the doctor should be able to use historical data to estimate the probabilities.

For a concrete example, we will consider a patient with a rash that matches a symptom of a rare disease affecting only 1 in 10 million individuals, but the symptom is always present in patients with that disease. However, the doctor knows that such a rash appears in 10% of patients with a more common condition affecting 1 in 50,000 individuals. What is the likelihood the patient has the rare disease?

We write $P(R) = 10^{-7}$ as the prior probability of the rare disease, $P(C) = 2 \times 10^{-5}$ as the prior probability of the more common illness, and $P(S)$ as the probability of the symptoms. From the question, we know $P(S|R) = 1$ (probability of the symptoms in someone with the rare disease is 1) and $P(S|C) = 0.1$ (probability of the symptoms in someone with the common condition is 0.1).

If the rare and common diseases are the only ways to get the symptoms then $P(S) = P(S|R)P(R) + P(S|C)P(C)$ (using *Rule 3*).

Then Bayes' Theorem (*Rule* 4, Eq. 1.32) tells us the probability of the rare disease given the symptoms is:

$$P(R|S) = \frac{P(S|R)P(R)}{P(S)} = \frac{P(S|R)P(R)}{P(S|R)P(R) + P(S|C)P(C)}$$
$$= \frac{1 \times 10^{-7}}{1 \times 10^{-7} + 0.1 \times 2 \times 10^{-5}}$$
$$= 0.0476.$$

That is, the patient is 20 times more likely to have the common disease even though the symptoms are 10 times more likely for a patient with the rare disease than for a patient with the common disease.

(iii) <u>A lie-detector test.</u>

Suppose that 90% of criminals fail a lie-detector test when responding "No" when asked if they committed their crime, but 1 in 100 innocent parties fail the same test because of nervousness in the interrogation room. In a city of 100,000, a person is randomly selected and fails the lie detector test when asked about a crime that could only have been committed by a resident. What is the probability that the person did in fact commit the crime?

In this example, we write $P(C) = \frac{1}{100{,}000}$ as the prior probability of a randomly selected person being the criminal, $P(I) = P(\tilde{C}) = 1 - P(C) = \frac{99{,}999}{100{,}000}$ as the prior probability a randomly selected person is innocent, and $P(F)$ as the probability that someone fails the lie detector test. From the question, we know $P(F|C) = 0.9$ and $P(F|I) = 0.01$.

We can evaluate $P(F) = P(F|C)P(C) + P(F|I)\,P(I)$ (from *Rule* 3) and calculate the probability that someone who fails the test is a criminal via Bayes' Theorem (*Rule 4*, Eq. 1.32):

$$P(C|F) = \frac{P(F|C)P(C)}{P(F)} = \frac{P(F|C)P(C)}{P(F|C)P(C) + P(F|I)\,P(I)}$$
$$= \frac{0.9 \times 10^{-5}}{0.9 \times 10^{-5} + 0.01 \times 0.99999}$$
$$= 0.0009.$$

That is, the probability is still less than 1 in 1000. In this example, where the prior probability is so low, some intuition into the result from Bayes' Theorem can be gained by noting that a criminal is 90 times more likely to fail the test than an innocent person so the posterior probability that someone is the criminal increases from the prior probability by a factor of approximately 90.