Introduction

Probability is one of the most important disciplines in all of the sciences. It is also one of the least well understood.

Probability is especially important in computer science—it arises in virtually every branch of the field. In algorithm design and game theory, for example, algorithms and strategies that make random choices at certain steps frequently outperform deterministic algorithms and strategies. In information theory and signal processing, an understanding of randomness is critical for filtering out noise and compressing data. In cryptography and digital rights management, probability is crucial for achieving security. The list of examples is long.

Given the impact that probability has on computer science, it seems strange that probability should be so misunderstood by so many. The trouble is that "commonsense" intuition is demonstrably unreliable when it comes to problems involving random events. As a consequence, many students develop a fear of probability. We've witnessed many graduate oral exams where a student will solve the most horrendous calculation, only to then be tripped up by the simplest probability question. Even some faculty will start squirming if you ask them a question that starts "What is the probability that. . . ?"

Our goal in the remaining chapters is to equip you with the tools that will enable you to solve basic problems involving probability easily and confidently.

Chapter 17 introduces the basic definitions and an elementary 4-step process that can be used to determine the probability that a specified event occurs. We illustrate the method on two famous problems where your intuition will probably fail you. The key concepts of conditional probability and independence are introduced, along with examples of their use, and regrettable misuse, in practice: the probability you have a disease given that a diagnostic test says you do, and the probability that a suspect is guilty given that his blood type matches the blood found at the scene of the crime.

Random variables provide a more quantitative way to measure random events, and we study them in Chapter 19. For example, instead of determining the probability that it will rain, we may want to determine *how much* or *how long* it is likely to rain. The fundamental concept of the *expected value* of a random variable is introduced and some of its key properties are developed.

Chapter 20 examines the probability that a random variable deviates significantly from its expected value. Probability of deviation provides the theoretical basis for estimation by sampling which is fundamental in science, engineering, and human affairs. It is also especially important in engineering practice, where things are generally fine if they are going as expected, and you would like to be assured that the probability of an unexpected event is very low.

A final chapter applies the previous probabilistic tools to solve problems involving more complex random processes. You will see why you will probably never get very far ahead at the casino and how two Stanford graduate students became billionaires by combining graph theory and probability theory to design a better search engine for the web.

17 Events and Probability Spaces

17.1 Let’s Make a Deal

In the September 9, 1990 issue of *Parade* magazine, columnist Marilyn vos Savant responded to this letter:

> *Suppose you’re on a game show, and you’re given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say number 1, and the host, who knows what’s behind the doors, opens another door, say number 3, which has a goat. He says to you, “Do you want to pick door number 2?” Is it to your advantage to switch your choice of doors?*
>
> Craig. F. Whitaker
> Columbia, MD

The letter describes a situation like one faced by contestants in the 1970’s game show *Let’s Make a Deal*, hosted by Monty Hall and Carol Merrill. Marilyn replied that the contestant should indeed switch. She explained that if the car was behind either of the two unpicked doors—which is twice as likely as the the car being behind the picked door—the contestant wins by switching. But she soon received a torrent of letters, many from mathematicians, telling her that she was wrong. The problem became known as the *Monty Hall Problem* and it generated thousands of hours of heated debate.

This incident highlights a fact about probability: the subject uncovers lots of examples where ordinary intuition leads to completely wrong conclusions. So until you’ve studied probabilities enough to have refined your intuition, a way to avoid errors is to fall back on a rigorous, systematic approach such as the Four Step Method that we will describe shortly. First, let’s make sure we really understand the setup for this problem. This is always a good thing to do when you are dealing with probability.

17.1.1 Clarifying the Problem

Craig’s original letter to Marilyn vos Savant is a bit vague, so we must make some assumptions in order to have any hope of modeling the game formally. For example, we will assume that:

1. The car is equally likely to be hidden behind each of the three doors.
2. The player is equally likely to pick each of the three doors, regardless of the car’s location.
3. After the player picks a door, the host *must* open a different door with a goat behind it and offer the player the choice of staying with the original door or switching.
4. If the host has a choice of which door to open, then he is equally likely to select each of them.

In making these assumptions, we’re reading a lot into Craig Whitaker’s letter. There are other plausible interpretations that lead to different answers. But let’s accept these assumptions for now and address the question, “What is the probability that a player who switches wins the car?”

17.2 The Four Step Method

Every probability problem involves some sort of randomized experiment, process, or game. And each such problem involves two distinct challenges:

1. How do we model the situation mathematically?
2. How do we solve the resulting mathematical problem?

In this section, we introduce a four step approach to questions of the form, “What is the probability that. . . ?” In this approach, we build a probabilistic model step by step, formalizing the original question in terms of that model. Remarkably, this structured approach provides simple solutions to many famously confusing problems. For example, as you’ll see, the four step method cuts through the confusion surrounding the Monty Hall problem like a Ginsu knife.

17.2.1 Step 1: Find the Sample Space

Our first objective is to identify all the possible outcomes of the experiment. A typical experiment involves several randomly-determined quantities. For example, the Monty Hall game involves three such quantities:

1. The door concealing the car.
2. The door initially chosen by the player.

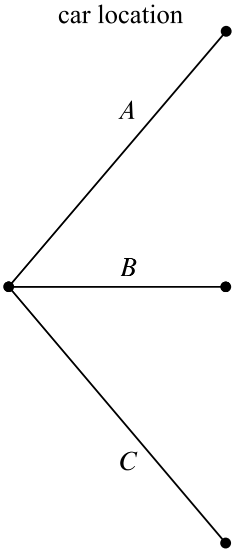


Figure 17.1 The first level in a tree diagram for the Monty Hall Problem. The branches correspond to the door behind which the car is located.

3. The door that the host opens to reveal a goat.

Every possible combination of these randomly-determined quantities is called an *outcome*. The set of all possible outcomes is called the *sample space* for the experiment.

A *tree diagram* is a graphical tool that can help us work through the four step approach when the number of outcomes is not too large or the problem is nicely structured. In particular, we can use a tree diagram to help understand the sample space of an experiment. The first randomly-determined quantity in our experiment is the door concealing the prize. We represent this as a tree with three branches, as shown in Figure 17.1. In this diagram, the doors are called A, B and C instead of 1, 2, and 3, because we'll be adding a lot of other numbers to the picture later.

For each possible location of the prize, the player could initially choose any of the three doors. We represent this in a second layer added to the tree. Then a third layer represents the possibilities of the final step when the host opens a door to reveal a goat, as shown in Figure 17.2.

Notice that the third layer reflects the fact that the host has either one choice or two, depending on the position of the car and the door initially selected by the player. For example, if the prize is behind door A and the player picks door B, then

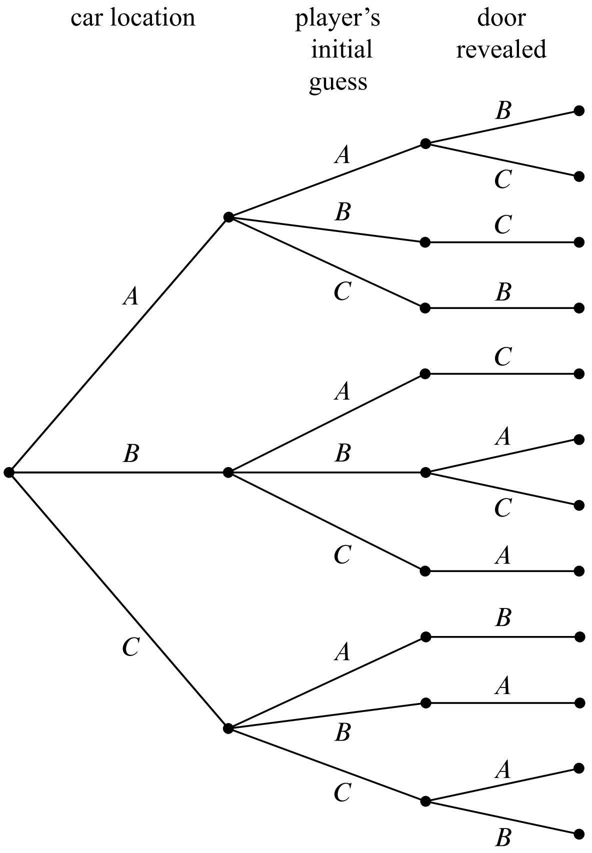


Figure 17.2 The full tree diagram for the Monty Hall Problem. The second level indicates the door initially chosen by the player. The third level indicates the door revealed by Monty Hall.

the host must open door C. However, if the prize is behind door A and the player picks door A, then the host could open either door B or door C.

Now let's relate this picture to the terms we introduced earlier: the leaves of the tree represent *outcomes* of the experiment, and the set of all leaves represents the *sample space*. Thus, for this experiment, the sample space consists of 12 outcomes. For reference, we've labeled each outcome in Figure 17.3 with a triple of doors indicating:

$$(\text{door concealing prize}, \ \text{door initially chosen}, \ \text{door opened to reveal a goat}).$$

In these terms, the sample space is the set

$$\mathcal{S} = \left\{ \begin{array}{l} (A,A,B),\ (A,A,C),\ (A,B,C),\ (A,C,B),\ (B,A,C),\ (B,B,A), \\ (B,B,C),\ (B,C,A),\ (C,A,B),\ (C,B,A),\ (C,C,A),\ (C,C,B) \end{array} \right\}$$

The tree diagram has a broader interpretation as well: we can regard the whole experiment as following a path from the root to a leaf, where the branch taken at each stage is "randomly" determined. Keep this interpretation in mind; we'll use it again later.

17.2.2 Step 2: Define Events of Interest

Our objective is to answer questions of the form "What is the probability that … ?", where, for example, the missing phrase might be "the player wins by switching," "the player initially picked the door concealing the prize," or "the prize is behind door C."

A set of outcomes is called an *event*. Each of the preceding phrases characterizes an event. For example, the event [prize is behind door C] refers to the set:

$$\{(C,A,B),(C,B,A),(C,C,A),(C,C,B)\},$$

and the event [prize is behind the door first picked by the player] is:

$$\{(A,A,B),(A,A,C),(B,B,A),(B,B,C),(C,C,A),(C,C,B)\}.$$

Here we're using square brackets around a property of outcomes as a notation for the event whose outcomes are the ones that satisfy the property.

What we're really after is the event [player wins by switching]:

$$\{(A,B,C),(A,C,B),(B,A,C),(B,C,A),(C,A,B),(C,B,A)\}. \tag{17.1}$$

The outcomes in this event are marked with checks in Figure 17.4.

Notice that exactly half of the outcomes are checked, meaning that the player wins by switching in half of all outcomes. You might be tempted to conclude that a player who switches wins with probability $1/2$. *This is wrong.* The reason is that these outcomes are not all equally likely, as we'll see shortly.

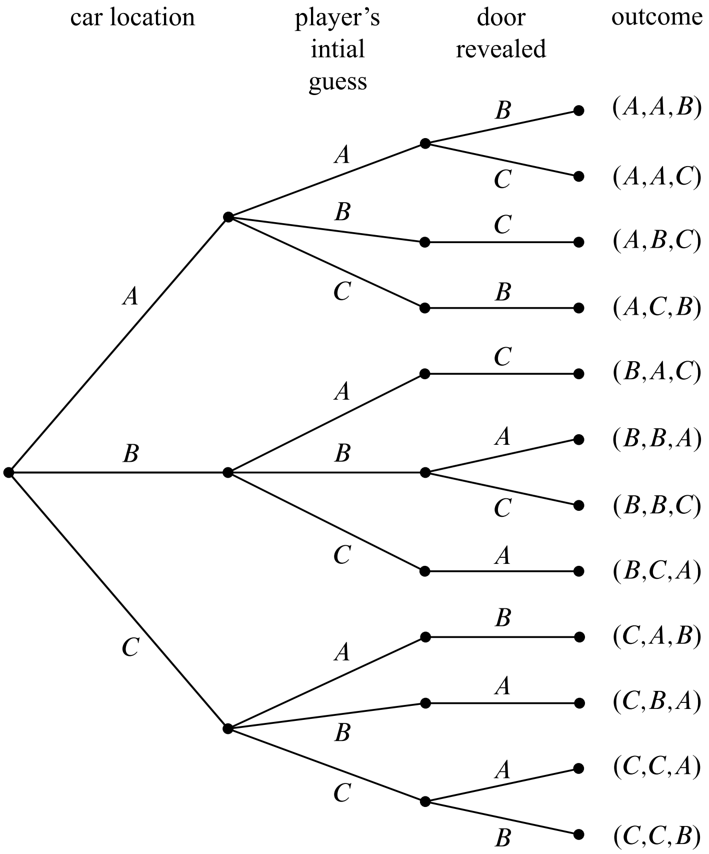


Figure 17.3 The tree diagram for the Monty Hall Problem with the outcomes labeled for each path from root to leaf. For example, outcome (A, A, B) corresponds to the car being behind door A, the player initially choosing door A, and Monty Hall revealing the goat behind door B.

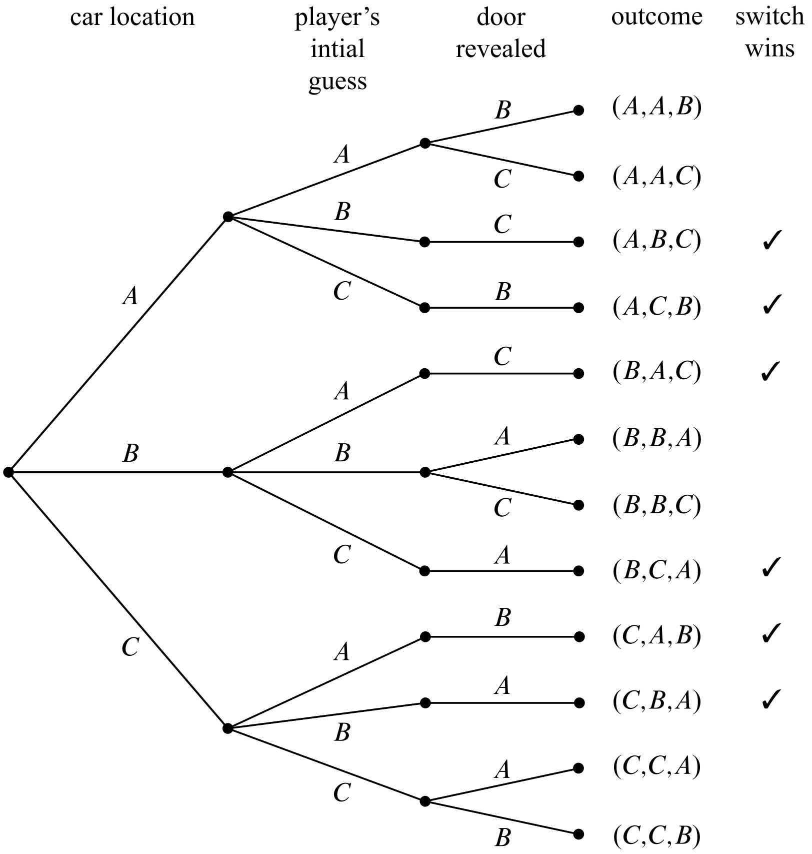


Figure 17.4 The tree diagram for the Monty Hall Problem, where the outcomes where the player wins by switching are denoted with a check mark.

17.2.3 Step 3: Determine Outcome Probabilities

So far we’ve enumerated all the possible outcomes of the experiment. Now we must start assessing the likelihood of those outcomes. In particular, the goal of this step is to assign each outcome a probability, indicating the fraction of the time this outcome is expected to occur. The sum of all the outcome probabilities must equal one, reflecting the fact that there always must be an outcome.

Ultimately, outcome probabilities are determined by the phenomenon we’re modeling and thus are not quantities that we can derive mathematically. However, mathematics can help us compute the probability of every outcome *based on fewer and more elementary modeling decisions.* In particular, we’ll break the task of determining outcome probabilities into two stages.

Step 3a: Assign Edge Probabilities

First, we record a probability on each *edge* of the tree diagram. These edge-probabilities are determined by the assumptions we made at the outset: that the prize is equally likely to be behind each door, that the player is equally likely to pick each door, and that the host is equally likely to reveal each goat, if he has a choice. Notice that when the host has no choice regarding which door to open, the single branch is assigned probability 1. For example, see Figure 17.5.

Step 3b: Compute Outcome Probabilities

Our next job is to convert edge probabilities into outcome probabilities. This is a purely mechanical process:

> calculate the probability of an outcome by multiplying the edge-probabilities on the path from the root to that outcome.

For example, the probability of the topmost outcome in Figure 17.5, (A, A, B), is

$$\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{18}. \tag{17.2}$$

We’ll examine the official justification for this rule in Section 18.4, but here’s an easy, intuitive justification: as the steps in an experiment progress randomly along a path from the root of the tree to a leaf, the probabilities on the edges indicate how likely the path is to proceed along each branch. For example, a path starting at the root in our example is equally likely to go down each of the three top-level branches.

How likely is such a path to arrive at the topmost outcome (A, A, B)? Well, there is a 1-in-3 chance that a path would follow the A-branch at the top level, a 1-in-3 chance it would continue along the A-branch at the second level, and 1-in-2 chance it would follow the B-branch at the third level. Thus, there is half of a one third of a one third chance, of arriving at the (A, A, B) leaf. That is, the chance is $1/3 \cdot 1/3 \cdot 1/2 = 1/18$—the same product (in reverse order) we arrived at in (17.2).

We have illustrated all of the outcome probabilities in Figure 17.5.

Specifying the probability of each outcome amounts to defining a function that maps each outcome to a probability. This function is usually called $\Pr[\cdot]$. In these terms, we've just determined that:

$$\begin{aligned}
\Pr[(A, A, B)] &= \frac{1}{18}, \\
\Pr[(A, A, C)] &= \frac{1}{18}, \\
\Pr[(A, B, C)] &= \frac{1}{9}, \\
&\text{etc.}
\end{aligned}$$

17.2.4 Step 4: Compute Event Probabilities

We now have a probability for each *outcome*, but we want to determine the probability of an *event*. The probability of an event E is denoted by $\Pr[E]$, and it is the sum of the probabilities of the outcomes in E. For example, the probability of the [switching wins] event (17.1) is

$$\begin{aligned}
&\Pr[\text{switching wins}] \\
&= \Pr[(A, B, C)] + \Pr[(A, C, B)] + \Pr[(B, A, C)] + \\
&\qquad \Pr[(B, C, A)] + \Pr[(C, A, B)] + \Pr[(C, B, A)] \\
&= \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} \\
&= \frac{2}{3}.
\end{aligned}$$

It seems Marilyn's answer is correct! A player who switches doors wins the car with probability $2/3$. In contrast, a player who stays with his or her original door wins with probability $1/3$, since staying wins if and only if switching loses.

We're done with the problem! We didn't need any appeals to intuition or ingenious analogies. In fact, no mathematics more difficult than adding and multiplying fractions was required. The only hard part was resisting the temptation to leap to an "intuitively obvious" answer.

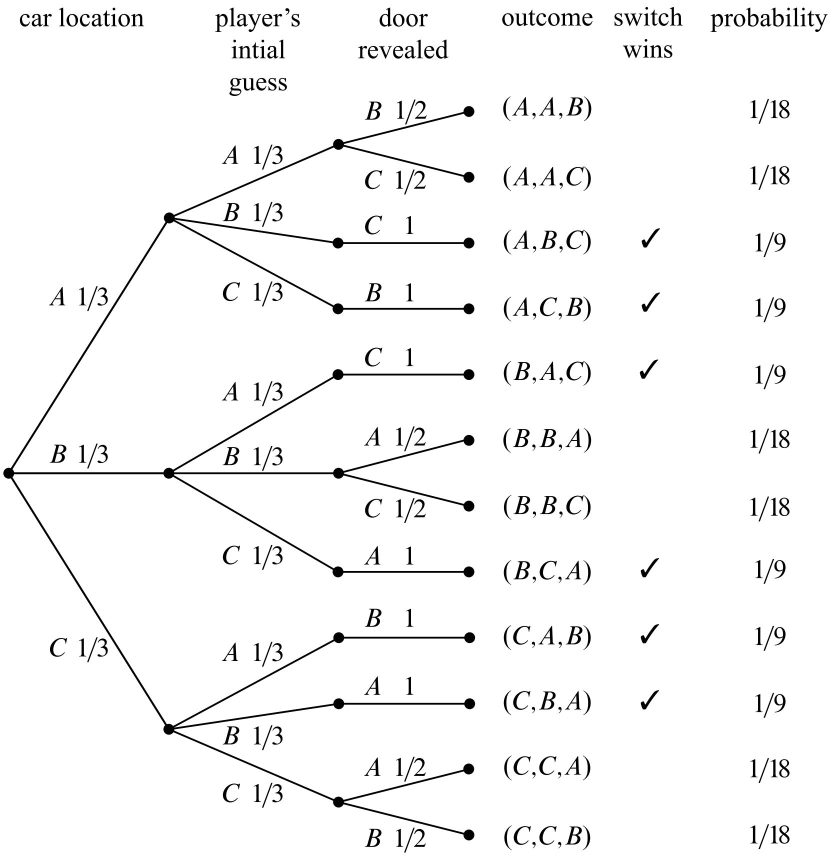


Figure 17.5 The tree diagram for the Monty Hall Problem where edge weights denote the probability of that branch being taken given that we are at the parent of that branch. For example, if the car is behind door A, then there is a 1/3 chance that the player's initial selection is door B. The rightmost column shows the outcome probabilities for the Monty Hall Problem. Each outcome probability is simply the product of the probabilities on the path from the root to the outcome leaf.

17.2.5 An Alternative Interpretation of the Monty Hall Problem

Was Marilyn really right? Our analysis indicates that she was. But a more accurate conclusion is that her answer is correct *provided we accept her interpretation of the question*. There is an equally plausible interpretation in which Marilyn's answer is wrong. Notice that Craig Whitaker's original letter does not say that the host is *required* to reveal a goat and offer the player the option to switch, merely that he *did* these things. In fact, on the *Let's Make a Deal* show, Monty Hall sometimes simply opened the door that the contestant picked initially. Therefore, if he wanted to, Monty could give the option of switching only to contestants who picked the correct door initially. In this case, switching never works!

17.3 Strange Dice

The four-step method is surprisingly powerful. Let's get some more practice with it. Imagine, if you will, the following scenario.

It's a typical Saturday night. You're at your favorite pub, contemplating the true meaning of infinite cardinalities, when a burly-looking biker plops down on the stool next to you. Just as you are about to get your mind around pow(pow($\mathbb{R}$)), biker dude slaps three strange-looking dice on the bar and challenges you to a \$100 wager. His rules are simple. Each player selects one die and rolls it once. The player with the lower value pays the other player \$100.

Naturally, you are skeptical, especially after you see that these are not ordinary dice. Each die has the usual six sides, but opposite sides have the same number on them, and the numbers on the dice are different, as shown in Figure 17.6.

Biker dude notices your hesitation, so he sweetens his offer: he will pay you \$105 if you roll the higher number, but you only need pay him \$100 if he rolls higher, *and* he will let you pick a die first, after which he will pick one of the other two. The sweetened deal sounds persuasive since it gives you a chance to pick what you think is the best die, so you decide you will play. But which of the dice should you choose? Die B is appealing because it has a 9, which is a sure winner if it comes up. Then again, die A has two fairly large numbers, and die C has an 8 and no really small values.

In the end, you choose die B because it has a 9, and then biker dude selects die A. Let's see what the probability is that you will win. (Of course, you probably should have done this before picking die B in the first place.) Not surprisingly, we will use the four-step method to compute this probability.

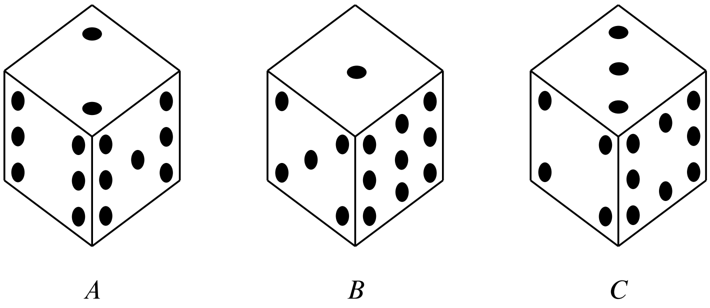


Figure 17.6 The strange dice. The number of pips on each concealed face is the same as the number on the opposite face. For example, when you roll die A, the probabilities of getting a 2, 6, or 7 are each $1/3$.

17.3.1 Die A versus Die B

Step 1: Find the sample space.
The tree diagram for this scenario is shown in Figure 17.7. In particular, the sample space for this experiment are the nine pairs of values that might be rolled with Die A and Die B:

For this experiment, the sample space is a set of nine outcomes:

$$\mathcal{S} = \{(2,1), (2,5), (2,9), (6,1), (6,5), (6,9), (7,1), (7,5), (7,9)\}.$$

Step 2: Define events of interest.
We are interested in the event that the number on die A is greater than the number on die B. This event is a set of five outcomes:

$$\{(2,1), (6,1), (6,5), (7,1), (7,5)\}.$$

These outcomes are marked A in the tree diagram in Figure 17.7.

Step 3: Determine outcome probabilities.
To find outcome probabilities, we first assign probabilities to edges in the tree diagram. Each number on each die comes up with probability $1/3$, regardless of the value of the other die. Therefore, we assign all edges probability $1/3$. The probability of an outcome is the product of the probabilities on the corresponding root-to-leaf path, which means that every outcome has probability $1/9$. These probabilities are recorded on the right side of the tree diagram in Figure 17.7.

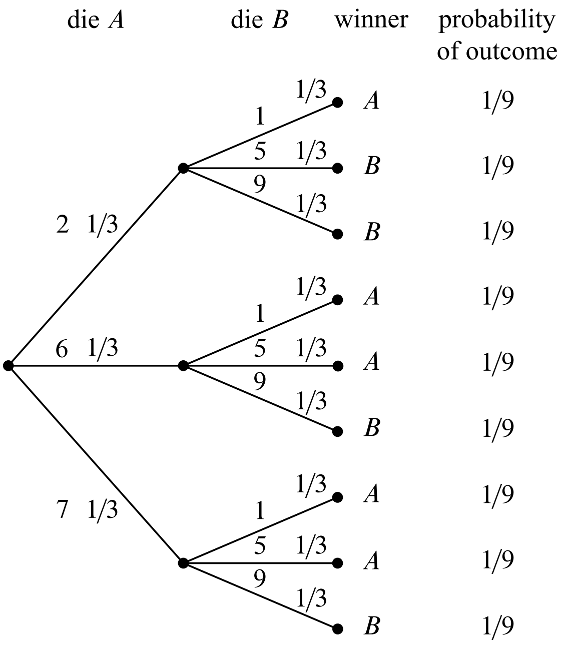


Figure 17.7 The tree diagram for one roll of die A versus die B. Die A wins with probability $5/9$.

Step 4: Compute event probabilities.
The probability of an event is the sum of the probabilities of the outcomes in that event. In this case, all the outcome probabilities are the same, so we say that the sample space is *uniform*. Computing event probabilities for uniform sample spaces is particularly easy since you just have to compute the number of outcomes in the event. In particular, for any event E in a uniform sample space $\mathcal{S}$,

$$\Pr[E] = \frac{|E|}{|\mathcal{S}|}. \tag{17.3}$$

In this case, E is the event that die A beats die B, so $|E| = 5$, $|\mathcal{S}| = 9$, and

$$\Pr[E] = 5/9.$$

This is bad news for you. Die A beats die B more than half the time and, not surprisingly, you just lost \$100.

Biker dude consoles you on your “bad luck” and, given that he’s a sensitive guy beneath all that leather, he offers to go double or nothing.[1] Given that your wallet only has \$25 in it, this sounds like a good plan. Plus, you figure that choosing die A will give *you* the advantage.

So you choose A, and then biker dude chooses C. Can you guess who is more likely to win? (Hint: it is generally not a good idea to gamble with someone you don’t know in a bar, especially when you are gambling with strange dice.)

17.3.2 Die A versus Die C

We can construct the tree diagram and outcome probabilities as before. The result is shown in Figure 17.8, and there is bad news again. Die C will beat die A with probability $5/9$, and you lose once again.

You now owe the biker dude \$200 and he asks for his money. You reply that you need to go to the bathroom.

17.3.3 Die B versus Die C

Being a sensitive guy, biker dude nods understandingly and offers yet another wager. This time, he’ll let you have die C. He’ll even let you raise the wager to \$200 so you can win your money back.

This is too good a deal to pass up. You know that die C is likely to beat die A and that die A is likely to beat die B, and so die C is *surely* the best. Whether biker

[1] *Double or nothing* is slang for doing another wager after you have lost the first. If you lose again, you will owe biker dude *double* what you owed him before. If you win, you will owe him *nothing*; in fact, since he should pay you \$210 if he loses, you would come out \$10 ahead.

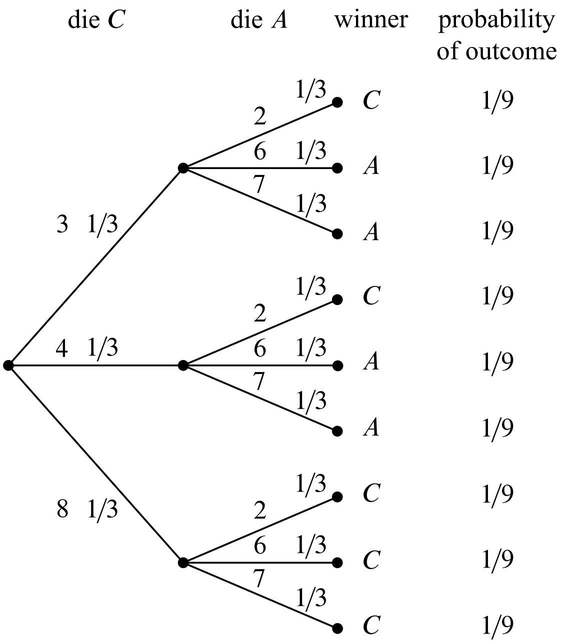


Figure 17.8 The tree diagram for one roll of die C versus die A. Die C wins with probability $5/9$.

dude picks A or B, the odds would be in your favor this time. Biker dude must really be a nice guy.

So you pick C, and then biker dude picks B. Wait—how come you haven't caught on yet and worked out the tree diagram before you took this bet? If you do it now, you'll see by the same reasoning as before that B beats C with probability $5/9$. But surely there is a mistake! How is it possible that

C beats A with probability $5/9$,

A beats B with probability $5/9$,

B beats C with probability $5/9$?

The problem is not with the math, but with your intuition. Since A will beat B more often than not, and B will beat C more often than not, it *seems* like A ought to beat C more often than not, that is, the "beats more often" relation ought to be *transitive*. But this intuitive idea is simply false: whatever die you pick, biker dude can pick one of the others and be likely to win. So picking first is actually a disadvantage, and as a result, you now owe biker dude \$400.

Just when you think matters can't get worse, biker dude offers you one final wager for \$1,000. This time, instead of rolling each die once, you will each roll your die twice, and your score is the sum of your rolls, and he will even let you pick your die second, that is, after he picks his. Biker dude chooses die B. Now you know that die A will beat die B with probability $5/9$ on one roll, so, jumping at this chance to get ahead, you agree to play, and you pick die A. After all, you figure that since a roll of die A beats a roll of die B more often that not, two rolls of die A are even more likely to beat two rolls of die B, right?

Wrong! (Did we mention that playing strange gambling games with strangers in a bar is a bad idea?)

17.3.4 Rolling Twice

If each player rolls twice, the tree diagram will have four levels and $3^4 = 81$ outcomes. This means that it will take a while to write down the entire tree diagram. But it's easy to write down the first two levels as in Figure 17.9(a) and then notice that the remaining two levels consist of nine identical copies of the tree in Figure 17.9(b).

The probability of each outcome is $(1/3)^4 = 1/81$ and so, once again, we have a uniform probability space. By equation (17.3), this means that the probability that A wins is the number of outcomes where A beats B divided by 81.

To compute the number of outcomes where A beats B, we observe that the two rolls of die A result in nine equally likely outcomes in a sample space $\mathcal{S}_A$ in which

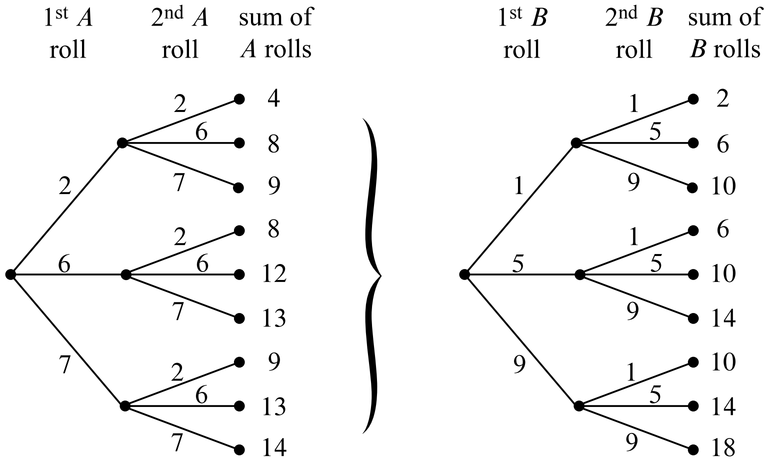


Figure 17.9 Parts of the tree diagram for die B versus die A where each die is rolled twice. The first two levels are shown in (a). The last two levels consist of nine copies of the tree in (b).

the two-roll sums take the values

$$(4, 8, 8, 9, 9, 12, 13, 13, 14).$$

Likewise, two rolls of die B result in nine equally likely outcomes in a sample space $\mathcal{S}_B$ in which the two-roll sums take the values

$$(2, 6, 6, 10, 10, 10, 14, 14, 18).$$

We can treat the outcome of rolling both dice twice as a pair $(x, y) \in \mathcal{S}_A \times \mathcal{S}_B$, where A wins iff the sum of the two A-rolls of outcome x is larger the sum of the two B-rolls of outcome y. If the A-sum is 4, there is only one y with a smaller B-sum, namely, when the B-sum is 2. If the A-sum is 8, there are three y's with a smaller B-sum, namely, when the B-sum is 2 or 6. Continuing the count in this way, the number of pairs (x, y) for which the A-sum is larger than the B-sum is

$$1 + 3 + 3 + 3 + 3 + 6 + 6 + 6 + 6 = 37.$$

A similar count shows that there are 42 pairs for which B-sum is larger than the A-sum, and there are two pairs where the sums are equal, namely, when they both equal 14. This means that A *loses* to B with probability $42/81 > 1/2$ and ties with probability $2/81$. Die A wins with probability only $37/81$.

How can it be that A is more likely than B to win with one roll, but B is more likely to win with two rolls? Well, why not? The only reason we'd think otherwise is our unreliable, untrained intuition. (Even the authors were surprised when they first learned about this, but at least they didn't lose \$1400 to biker dude.) In fact, the die strength reverses no matter which two die we picked. So for one roll,

$$A \succ B \succ C \succ A,$$

but for two rolls,

$$A \prec B \prec C \prec A,$$

where we have used the symbols $\succ$ and $\prec$ to denote which die is more likely to result in the larger value.

The weird behavior of the three strange dice above generalizes in a remarkable way: there are arbitrarily large sets of dice which will beat each other in any desired pattern according to how many times the dice are rolled.[2]

17.4 The Birthday Principle

There are 95 students in a class. What is the probability that some birthday is shared by two people? Comparing 95 students to the 365 possible birthdays, you might guess the probability lies somewhere around $1/4$—but you'd be wrong: the probability that there will be two people in the class with matching birthdays is actually more than 0.9999.

To work this out, we'll assume that the probability that a randomly chosen student has a given birthday is $1/d$. We'll also assume that a class is composed of n randomly and independently selected students. Of course $d = 365$ and $n = 95$ in this case, but we're interested in working things out in general. These randomness assumptions are not really true, since more babies are born at certain times of year, and students' class selections are typically not independent of each other, but simplifying in this way gives us a start on analyzing the problem. More importantly, these assumptions are justifiable in important computer science applications of birthday matching. For example, birthday matching is a good model for collisions between items randomly inserted into a hash table. So we won't worry about things like spring procreation preferences that make January birthdays more common, or about twins' preferences to take classes together (or not).

[2] **TBA - Reference Ron Graham paper.**

17.4.1 Exact Formula for Match Probability

There are d^n sequences of n birthdays, and under our assumptions, these are equally likely. There are $d(d-1)(d-2)\cdots(d-(n-1))$ length n sequences of distinct birthdays. That means the probability that everyone has a different birthday is:

$$\begin{aligned}
&\frac{d(d-1)(d-2)\cdots(d-(n-1))}{d^n} \\
&= \frac{d}{d}\cdot\frac{d-1}{d}\cdot\frac{d-2}{d}\cdots\frac{d-(n-1)}{d} && (17.4)\\
&= \left(1-\frac{0}{d}\right)\left(1-\frac{1}{d}\right)\left(1-\frac{2}{d}\right)\cdots\left(1-\frac{n-1}{d}\right) && (17.5)
\end{aligned}$$

Now we simplify (17.5) using the fact that $1-x < e^{-x}$ for all $x > 0$. This follows by truncating the Taylor series $e^{-x} = 1 - x + x^2/2! - x^3/3! + \cdots$. The approximation $e^{-x} \approx 1 - x$ is pretty accurate when x is small.

$$\begin{aligned}
&\left(1-\frac{0}{d}\right)\left(1-\frac{1}{d}\right)\left(1-\frac{2}{d}\right)\cdots\left(1-\frac{n-1}{d}\right) \\
&< e^0 \cdot e^{-1/d} \cdot e^{-2/d} \cdots e^{-(n-1)/d} && (17.6)\\
&= e^{-\left(\sum_{i=1}^{n-1} i/d\right)} \\
&= e^{-(n(n-1)/(2d))}. && (17.7)
\end{aligned}$$

For $n = 95$ and $d = 365$, the value of (17.7) is less than $1/200,000$, which means the probability of having some pair of matching birthdays actually is more than $1 - 1/200,000 > 0.99999$. So it would be pretty astonishing if there were no pair of students in the class with matching birthdays.

For $d \le n^2/2$, the probability of no match turns out to be asymptotically equal to the upper bound (17.7). For $d = n^2/2$ in particular, the probability of no match is asymptotically equal to $1/e$. This leads to a rule of thumb which is useful in many contexts in computer science:

The Birthday Principle

If there are d days in a year and $\sqrt{2d}$ people in a room, then the probability that two share a birthday is about $1 - 1/e \approx 0.632$.

For example, the Birthday Principle says that if you have $\sqrt{2 \cdot 365} \approx 27$ people in a room, then the probability that two share a birthday is about 0.632. The actual probability is about 0.626, so the approximation is quite good.

Among other applications, it implies that to use a hash function that maps n items into a hash table of size d, you can expect many collisions if n^2 is more than a small fraction of d. The Birthday Principle also famously comes into play as the basis of "birthday attacks" that crack certain cryptographic systems.

17.5 Set Theory and Probability

Let's abstract what we've just done into a general mathematical definition of sample spaces and probability.

17.5.1 Probability Spaces

Definition 17.5.1. A countable *sample space* $\mathcal{S}$ is a nonempty countable set.[3] An element $\omega \in \mathcal{S}$ is called an *outcome*. A subset of $\mathcal{S}$ is called an *event*.

Definition 17.5.2. A *probability function* on a sample space $\mathcal{S}$ is a total function $\Pr : \mathcal{S} \to \mathbb{R}$ such that

- $\Pr[\omega] \geq 0$ for all $\omega \in \mathcal{S}$, and
- $\sum_{\omega \in \mathcal{S}} \Pr[\omega] = 1$.

A sample space together with a probability function is called a *probability space*. For any event $E \subseteq \mathcal{S}$, the *probability of* E is defined to be the sum of the probabilities of the outcomes in E:

$$\Pr[E] ::= \sum_{\omega \in E} \Pr[\omega].$$

In the previous examples there were only finitely many possible outcomes, but we'll quickly come to examples that have a countably infinite number of outcomes.

The study of probability is closely tied to set theory because any set can be a sample space and any subset can be an event. General probability theory deals with uncountable sets like the set of real numbers, but we won't need these, and sticking to countable sets lets us define the probability of events using sums instead of integrals. It also lets us avoid some distracting technical problems in set theory like the Banach-Tarski "paradox" mentioned in Chapter 8.

[3] Yes, sample spaces can be infinite. If you did not read Chapter 8, don't worry—*countable* just means that you can list the elements of the sample space as $\omega_0, \omega_1, \omega_2, \ldots$.

17.5.2 Probability Rules from Set Theory

Most of the rules and identities that we have developed for finite sets extend very naturally to probability.

An immediate consequence of the definition of event probability is that for *disjoint* events E and F,

$$\Pr[E \cup F] = \Pr[E] + \Pr[F].$$

This generalizes to a countable number of events:

Rule 17.5.3 (Sum Rule)**.** *If $E_0, E_1, \ldots, E_n, \ldots$ are pairwise disjoint events, then*

$$\Pr\left[\bigcup_{n \in \mathbb{N}} E_n\right] = \sum_{n \in \mathbb{N}} \Pr[E_n].$$

The Sum Rule lets us analyze a complicated event by breaking it down into simpler cases. For example, if the probability that a randomly chosen MIT student is native to the United States is 60%, to Canada is 5%, and to Mexico is 5%, then the probability that a random MIT student is native to one of these three countries is 70%.

Another consequence of the Sum Rule is that $\Pr[A] + \Pr[\overline{A}] = 1$, which follows because $\Pr[\mathcal{S}] = 1$ and $\mathcal{S}$ is the union of the disjoint sets A and $\overline{A}$. This equation often comes up in the form:

$$\Pr[\overline{A}] = 1 - \Pr[A]. \qquad \text{(Complement Rule)}$$

Sometimes the easiest way to compute the probability of an event is to compute the probability of its complement and then apply this formula.

Some further basic facts about probability parallel facts about cardinalities of finite sets. In particular:

$$\begin{aligned}
&\Pr[B - A] = \Pr[B] - \Pr[A \cap B], && \text{(Difference Rule)}\\
&\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B], && \text{(Inclusion-Exclusion)}\\
&\Pr[A \cup B] \le \Pr[A] + \Pr[B], && \text{(Boole's Inequality)}\\
&\text{If } A \subseteq B, \text{ then } \Pr[A] \le \Pr[B]. && \text{(Monotonicity Rule)}
\end{aligned}$$

The Difference Rule follows from the Sum Rule because B is the union of the disjoint sets $B - A$ and $A \cap B$. Inclusion-Exclusion then follows from the Sum and Difference Rules, because $A \cup B$ is the union of the disjoint sets A and $B - A$. Boole's inequality is an immediate consequence of Inclusion-Exclusion since probabilities are nonnegative. Monotonicity follows from the definition of event probability and the fact that outcome probabilities are nonnegative.

The two-event Inclusion-Exclusion equation above generalizes to any finite set of events in the same way as the corresponding Inclusion-Exclusion rule for n sets. Boole’s inequality also generalizes to both finite and countably infinite sets of events:

Rule 17.5.4 (Union Bound)**.**

$$\Pr[E_1 \cup \cdots \cup E_n \cup \cdots] \leq \Pr[E_1] + \cdots + \Pr[E_n] + \cdots . \tag{17.8}$$

The Union Bound is useful in many calculations. For example, suppose that E_i is the event that the i-th critical component among n components in a spacecraft fails. Then $E_1 \cup \cdots \cup E_n$ is the event that *some* critical component fails. If $\sum_{i=1}^{n} \Pr[E_i]$ is small, then the Union Bound can provide a reassuringly small upper bound on this overall probability of critical failure.

17.5.3 Uniform Probability Spaces

Definition 17.5.5. A finite probability space $\mathcal{S}$ is said to be *uniform* if $\Pr[\omega]$ is the same for every outcome $\omega \in \mathcal{S}$.

As we saw in the strange dice problem, uniform sample spaces are particularly easy to work with. That’s because for any event $E \subseteq \mathcal{S}$,

$$\Pr[E] = \frac{|E|}{|\mathcal{S}|}. \tag{17.9}$$

This means that once we know the cardinality of E and $\mathcal{S}$, we can immediately obtain $\Pr[E]$. That’s great news because we developed lots of tools for computing the cardinality of a set in Part III.

For example, suppose that you select five cards at random from a standard deck of 52 cards. What is the probability of having a full house? Normally, this question would take some effort to answer. But from the analysis in Section 15.7.2, we know that

$$|\mathcal{S}| = \binom{52}{5}$$

and

$$|E| = 13 \cdot \binom{4}{3} \cdot 12 \cdot \binom{4}{2}$$

where E is the event that we have a full house. Since every five-card hand is equally

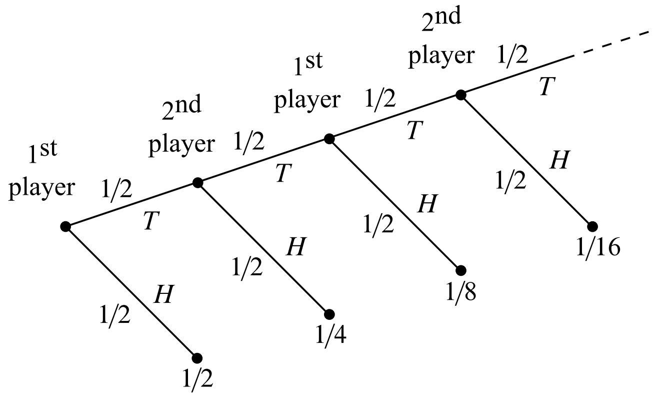


Figure 17.10 The tree diagram for the game where players take turns flipping a fair coin. The first player to flip heads wins.

likely, we can apply equation (17.9) to find that

$$\begin{aligned}\Pr[E] &= \frac{13 \cdot 12 \cdot \binom{4}{3} \cdot \binom{4}{2}}{\binom{52}{5}} \\ &= \frac{13 \cdot 12 \cdot 4 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} = \frac{18}{12495} \\ &\approx \frac{1}{694}.\end{aligned}$$

17.5.4 Infinite Probability Spaces

Infinite probability spaces are fairly common. For example, two players take turns flipping a fair coin. Whoever flips heads first is declared the winner. What is the probability that the first player wins? A tree diagram for this problem is shown in Figure 17.10.

The event that the first player wins contains an infinite number of outcomes, but we can still sum their probabilities:

$$\begin{aligned}\Pr[\text{first player wins}] &= \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \frac{1}{128} + \cdots \\ &= \frac{1}{2}\sum_{n=0}^{\infty}\left(\frac{1}{4}\right)^n \\ &= \frac{1}{2}\left(\frac{1}{1-1/4}\right) = \frac{2}{3}.\end{aligned}$$

Similarly, we can compute the probability that the second player wins:

$$\Pr[\text{second player wins}] = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \frac{1}{256} + \cdots = \frac{1}{3}.$$

In this case, the sample space is the infinite set

$$\mathcal{S} ::= \{\, \mathtt{T}^n\mathtt{H} \mid n \in \mathbb{N} \,\},$$

where $\mathtt{T}^n$ stands for a length n string of $\mathtt{T}$’s. The probability function is

$$\Pr[\mathtt{T}^n\mathtt{H}] ::= \frac{1}{2^{n+1}}.$$

To verify that this is a probability space, we just have to check that all the probabilities are nonnegative and that they sum to 1. The given probabilities are all nonnegative, and applying the formula for the sum of a geometric series, we find that

$$\sum_{n \in \mathbb{N}} \Pr[\mathtt{T}^n\mathtt{H}] = \sum_{n \in \mathbb{N}} \frac{1}{2^{n+1}} = 1.$$

Notice that this model does not have an outcome corresponding to the possibility that both players keep flipping tails forever. (In the diagram, flipping forever corresponds to following the infinite path in the tree without ever reaching a leaf/outcome.) If leaving this possibility out of the model bothers you, you’re welcome to fix it by adding another outcome ω_{forever} to indicate that that’s what happened. Of course since the probabililities of the other outcomes already sum to 1, you have to define the probability of ω_{forever} to be 0. Now outcomes with probability zero will have no impact on our calculations, so there’s no harm in adding it in if it makes you happier. On the other hand, in countable probability spaces it isn’t necessary to have outcomes with probability zero, and we will generally ignore them.

17.6 References

[19], [26], [30], [34], [38], [39] [43], [42], [51]

Problems for Section 17.2

Practice Problems

Problem 17.1.
Let B be the number of heads that come up on $2n$ independent tosses of a fair coin.

(a) $\Pr[B = n]$ is asymptotically equal to one of the expressions given below. Explain which one.

1. $\frac{1}{\sqrt{2\pi n}}$
2. $\frac{2}{\sqrt{\pi n}}$
3. $\frac{1}{\sqrt{\pi n}}$
4. $\sqrt{\frac{2}{\pi n}}$

Exam Problems

Problem 17.2. **(a)** What's the probability that 0 doesn't appear among k digits chosen independently and uniformly at random?

(b) A box contains 90 good and 10 defective screws. What's the probability that if we pick 10 screws from the box, none will be defective?

(c) First one digit is chosen uniformly at random from $\{1, 2, 3, 4, 5\}$ and is removed from the set; then a second digit is chosen uniformly at random from the remaining digits. What is the probability that an odd digit is picked the second time?

(d) Suppose that you *randomly* permute the digits $1, 2, \cdots, n$, that is, you select a permutation uniformly at random. What is the probability the digit k ends up in the ith position after the permutation?

(e) A fair coin is flipped n times. What's the probability that all the heads occur at the end of the sequence? (If no heads occur, then "all the heads are at the end of the sequence" is vacuously true.)

Class Problems

Problem 17.3.
The New York Yankees and the Boston Red Sox are playing a two-out-of-three

series. In other words, they play until one team has won two games. Then that team is declared the overall winner and the series ends. Assume that the Red Sox win each game with probability $3/5$, regardless of the outcomes of previous games.

Answer the questions below using the four step method. You can use the same tree diagram for all three problems.

(a) What is the probability that a total of 3 games are played?

(b) What is the probability that the winner of the series loses the first game?

(c) What is the probability that the *correct* team wins the series?

Problem 17.4.

To determine which of two people gets a prize, a coin is flipped twice. If the flips are a Head and then a Tail, the first player wins. If the flips are a Tail and then a Head, the second player wins. However, if both coins land the same way, the flips don't count and the whole process starts over.

Assume that on each flip, a Head comes up with probability p, regardless of what happened on other flips. Use the four step method to find a simple formula for the probability that the first player wins. What is the probability that neither player wins?

Hint: The tree diagram and sample space are infinite, so you're not going to finish drawing the tree. Try drawing only enough to see a pattern. Summing all the winning outcome probabilities directly is cumbersome. However, a neat trick solves this problem—and many others. Let s be the sum of all winning outcome probabilities in the whole tree. Notice that *you can write the sum of all the winning probabilities in certain subtrees as a function of* s. Use this observation to write an equation in s and then solve.

Homework Problems

Problem 17.5.

Let's see what happens when *Let's Make a Deal* is played with **four** doors. A prize is hidden behind one of the four doors. Then the contestant picks a door. Next, the host opens an unpicked door that has no prize behind it. The contestant is allowed to stick with their original door or to switch to one of the two unopened, unpicked doors. The contestant wins if their final choice is the door hiding the prize.

Let's make the same assumptions as in the original problem:

1. The prize is equally likely to be behind each door.

2. The contestant is equally likely to pick each door initially, regardless of the prize's location.

3. The host is equally likely to reveal each door that does not conceal the prize and was not selected by the player.

Use The Four Step Method to find the following probabilities. The tree diagram may become awkwardly large, in which case just draw enough of it to make its structure clear.

(a) Contestant Stu, a sanitation engineer from Trenton, New Jersey, stays with his original door. What is the probability that Stu wins the prize?

(b) Contestant Zelda, an alien abduction researcher from Helena, Montana, switches to one of the remaining two doors with equal probability. What is the probability that Zelda wins the prize?

Now let's revise our assumptions about how contestants choose doors. Say the doors are labeled A, B, C, and D. Suppose that Carol always opens the *earliest* door possible (the door whose label is earliest in the alphabet) with the restriction that she can neither reveal the prize nor open the door that the player picked.

This gives contestant Mergatroid—an engineering student from Cambridge, MA—just a little more information about the location of the prize. Suppose that Mergatroid always switches to the earliest door, excluding his initial pick and the one Carol opened.

(c) What is the probability that Mergatroid wins the prize?

Problem 17.6.
There were n Immortal Warriors born into our world, but in the end there can be *only one*. The Immortals' original plan was to stalk the world for centuries, dueling one another with ancient swords in dramatic landscapes until only one survivor remained. However, after a thought-provoking discussion probability, they opt to give the following protocol a try:

(i) The Immortals forge a coin that comes up heads with probability p.

(ii) Each Immortal flips the coin once.

(iii) If *exactly one* Immortal flips heads, then they are declared The One. Otherwise, the protocol is declared a failure, and they all go back to hacking each other up with swords.

One of the Immortals (Kurgan from the Russian steppe) argues that as n grows large, the probability that this protocol succeeds must tend to zero. Another (McLeod from the Scottish highlands) argues that this need not be the case, provided p is chosen carefully.

(a) A natural sample space to use to model this problem is $\{H, T\}^n$ of length-n sequences of H and T's, where the successive H's and T's in an outcome correspond to the Head or Tail flipped on each one of the n successive flips. Explain how a tree diagram approach leads to assigning a probability to each outcome that depends only on p, n and the number h of H's in the outcome.

(b) What is the probability that the experiment succeeds as a function of p and n?

(c) How should p, the bias of the coin, be chosen in order to maximize the probability that the experiment succeeds?

(d) What is the probability of success if p is chosen in this way? What quantity does this approach when n, the number of Immortal Warriors, grows large?

Problem 17.7.
We play a game with a deck of 52 regular playing cards, of which 26 are red and 26 are black. I randomly shuffle the cards and place the deck face down on a table. You have the option of "taking" or "skipping" the top card. If you skip the top card, then that card is revealed and we continue playing with the remaining deck. If you take the top card, then the game ends; you win if the card you took was revealed to be black, and you lose if it was red. If we get to a point where there is only one card left in the deck, you must take it. Prove that you have no better strategy than to take the top card—which means your probability of winning is 1/2.

Hint: Prove by induction the more general claim that for a randomly shuffled deck of n cards that are red or black—not necessarily with the same number of red cards and black cards—there is no better strategy than taking the top card.

Problems for Section 17.5

Class Problems

Problem 17.8.
Suppose there is a system with n components, and we know from past experience that any particular component will fail in a given year with probability p. That is,

letting F_i be the event that the ith component fails within one year, we have

$$\Pr[F_i] = p$$

for $1 \leq i \leq n$. The *system* will fail if *any one* of its components fails. What can we say about the probability that the system will fail within one year?

Let F be the event that the system fails within one year. Without any additional assumptions, we can't get an exact answer for $\Pr[F]$. However, we can give useful upper and lower bounds, namely,

$$p \leq \Pr[F] \leq np. \tag{17.10}$$

We may as well assume $p < 1/n$, since the upper bound is trivial otherwise. For example, if $n = 100$ and $p = 10^{-5}$, we conclude that there is at most one chance in 1000 of system failure within a year and at least one chance in 100,000.

Let's model this situation with the sample space $\mathcal{S} ::= \text{pow}([1..n])$ whose outcomes are subsets of positive integers $\leq n$, where $s \in \mathcal{S}$ corresponds to the indices of exactly those components that fail within one year. For example, $\{2, 5\}$ is the outcome that the second and fifth components failed within a year and none of the other components failed. So the outcome that the system did not fail corresponds to the empty set $\emptyset$.

(a) Show that the probability that the system fails could be as small as p by describing appropriate probabilities for the outcomes. Make sure to verify that the sum of your outcome probabilities is 1.

(b) Show that the probability that the system fails could actually be as large as np by describing appropriate probabilities for the outcomes. Make sure to verify that the sum of your outcome probabilities is 1.

(c) Prove inequality (17.10).

Problem 17.9.

Here are some handy rules for reasoning about probabilities that all follow directly from the Disjoint Sum Rule. Prove them.

$$\begin{aligned}
\Pr[A - B] &= \Pr[A] - \Pr[A \cap B] && \text{(Difference Rule)}\\
\Pr[\overline{A}] &= 1 - \Pr[A] && \text{(Complement Rule)}\\
\Pr[A \cup B] &= \Pr[A] + \Pr[B] - \Pr[A \cap B] && \text{(Inclusion-Exclusion)}\\
\Pr[A \cup B] &\leq \Pr[A] + \Pr[B] && \text{(2-event Union Bound)}\\
A \subseteq B \text{ IMPLIES } \Pr[A] &\leq \Pr[B] && \text{(Monotonicity)}
\end{aligned}$$

Homework Problems

Problem 17.10.

Prove the following probabilistic inequality, referred to as the *Union Bound*.

Let $A_1, A_2, \ldots, A_n, \ldots$ be events. Then

$$\Pr\left[\bigcup_{n\in\mathbb{N}} A_n\right] \le \sum_{n\in\mathbb{N}} \Pr[A_n].$$

Hint: Replace the A_n’s by pairwise disjoint events and use the Sum Rule.

Problem 17.11.

The results of a round robin tournament in which every two people play each other and one of them wins can be modelled a *tournament digraph*—a digraph with exactly one edge between each pair of distinct vertices, but we’ll continue to use the language of players beating each other.

An n-player tournament is k*-neutral* for some $k \in [0, n)$, when, for every set of k players, there is another player who beats them all. For example, being 1-neutral is the same as not having a “best” player who beats everyone else.

This problem shows that for any fixed k, if n is large enough, there will be a k-neutral tournament of n players. We will do this by reformulating the question in terms of probabilities. In particular, for any fixed n, we assign probabilities to each n-vertex tournament digraph by choosing a direction for the edge between any two vertices, independently and with equal probability for each edge.

(a) For any set S of k players, let B_S be the event that no contestant beats everyone in S. Express $\Pr[B_S]$ in terms of n and k.

(b) Let Q_k be the event that the tournament digraph is *not* k-neutral. Prove that

$$\Pr[Q_k] \le \binom{n}{k}\alpha^{n-k},$$

where $\alpha ::= 1 - (1/2)^k$.

Hint: Let S range over the size-k subsets of players, so

$$Q_k = \bigcup_S B_S\,.$$

Use Boole’s inequality.

(c) Conclude that if n is large enough (relative to k), then $\Pr[Q_k] < 1$.

(d) Explain why the previous result implies that for every integer k, there is an n-player k-neutral tournament (for a large enough $n \in \mathbb{N}$).

Homework Problems

Problem 17.12.

Suppose you repeatedly flip a fair coin until three consecutive flips match the pattern HHT or the pattern TTH occurs. What is the probability you will see HHT first? Define a suitable probability space that models the coin flipping and use it to explain your answer.

Hint: Symmetry between Heads and Tails.

18 Conditional Probability

18.1 Monty Hall Confusion

Remember how we said that the Monty Hall problem confused even professional mathematicians? Based on the work we did with tree diagrams, this may seem surprising—the conclusion we reached followed routinely and logically. How could this problem be so confusing to so many people?

Well, one flawed argument goes as follows: let's say the contestant picks door A. And suppose that Carol, Monty's assistant, opens door B and shows us a goat. Let's use the tree diagram 17.3 from Chapter 17 to capture this situation. There are exactly three outcomes where contestant chooses door A, and there is a goat behind door B:

$$(A, A, B),\ (A, A, C),\ (C, A, B). \tag{18.1}$$

These outcomes have respective probabilities 1/18, 1/18, 1/9.

Among those outcomes, switching doors wins only on the last outcome (C, A, B). The other two outcomes *together* have the *same* 1/9 probability as the last one So in this situation, the probability that we win by switching is the *same* as the probability that we lose. In other words, in this situation, switching isn't any better than sticking!

Something has gone wrong here, since we know that the actual probability of winning by switching in 2/3. The mistaken conclusion that sticking or switching are equally good strategies comes from a common blunder in reasoning about how probabilities change given some information about what happened. We have asked for the probability that one event, [win by switching], happens, *given* that another event, [pick A AND goat at B], happens. We use the notation

$$\Pr\big[[\text{win by switching}] \mid [\text{pick A AND goat at B}]\big]$$

for this probability which, by the reasoning above, equals 1/2.

18.1.1 Behind the Curtain

A "given" condition is essentially an instruction to focus on only some of the possible outcomes. Formally, we're defining a new sample space consisting only of some of the outcomes. In this particular example, we're given that the player chooses door A and that there is a goat behind B. Our new sample space therefore consists solely of the three outcomes listed in (18.1). In the opening of Section 18.1, we calculated the conditional probability of winning by switching given that one of these outcome happened, by weighing the 1/9 probability of the win-by-switching outcome (C, A, B) against the $1/18 + 1/18 + 1/9$ probability of the three outcomes in the new sample space.

$$\begin{aligned}
&\Pr\big[[\text{win by switching}] \mid [\text{pick A AND goat at B}]\big] \\
&= \Pr\big[(C, A, B) \mid \{(C, A, B), (A, A, B), (A, A, C)\}\big] + \\
&\qquad \frac{\Pr[(C, A, B)]}{\Pr[\{(C, A, B), (A, A, B), (A, A, C)\}]} \\
&= \frac{1/9}{1/18 + 1/18 + 1/9} = \frac{1}{2}.
\end{aligned}$$

There is nothing wrong with this calculation. So how come it leads to an incorrect conclusion about whether to stick or switch? The answer is that this was the wrong thing to calculate, as we'll explain in the next section.

18.2 Definition and Notation

The expression $\Pr\big[X \mid Y\big]$ denotes the probability of event X, given that event Y happens. In the example above, event X is the event of winning on a switch, and event Y is the event that a goat is behind door B and the contestant chose door A. We calculated $\Pr\big[X \mid Y\big]$ using a formula which serves as the definition of conditional probability:

Definition 18.2.1. Let X and Y be events where Y has nonzero probability. Then

$$\Pr\big[X \mid Y\big] ::= \frac{\Pr[X \cap Y]}{\Pr[Y]}.$$

The conditional probability $\Pr\big[X \mid Y\big]$ is undefined when the probability of event Y is zero. To avoid cluttering up statements with uninteresting hypotheses that conditioning events like Y have nonzero probability, we will make an implicit assumption from now on that all such events have nonzero probability.

Pure probability is often counterintuitive, but conditional probability can be even worse. Conditioning can subtly alter probabilities and produce unexpected results in randomized algorithms and computer systems as well as in betting games. But Definition 18.2.1 is very simple and causes no trouble—provided it is properly applied.

18.2.1 What went wrong

So if everything in the opening Section 18.1 is mathematically sound, why does it seem to contradict the results that we established in Chapter 17? The problem is a common one: *we chose the wrong condition.* In our initial description of the scenario, we learned the location of the goat when Carol opened door B. But when we defined our condition as "the contestant opens A and the goat is behind B," we included the outcome (A, A, C) in which Carol opens door C! The correct conditional probability should have been "what are the odds of winning by switching given the contestant chooses door A and Carol opens door B." By choosing a condition that did not reflect everything known. we inadvertently included an extraneous outcome in our calculation. With the correct conditioning, we still win by switching 1/9 of the time, but the smaller set of known outcomes has smaller total probability:

$$\Pr[\{(A,A,B),(C,A,B)\}] = \frac{1}{18} + \frac{1}{9} = \frac{3}{18}.$$

The conditional probability would then be:

$$\begin{aligned}
&\Pr\big[[\text{win by switching}] \mid [\text{pick A AND Carol opens B}]\big] \\
&= \Pr\big[(C,A,B) \mid \{(C,A,B),(A,A,B)\}\big] + \frac{\Pr[(C,A,B)]}{\Pr[\{(C,A,B),(A,A,B)\}]} \\
&= \frac{1/9}{1/9 + 1/18} = \frac{2}{3},
\end{aligned}$$

which is exactly what we already deduced from the tree diagram 17.2 in Section 17.2.

The O. J. Simpson Trial

In an opinion article in the *New York Times*, Steven Strogatz points to the O. J. Simpson trial as an example of poor choice of conditions. O. J. Simpson was a retired football player who was accused, and later acquitted, of the murder of his wife, Nicole Brown Simpson. The trial was widely publicized and called the “trial of the century.” Racial tensions, allegations of police misconduct, and new-at-the-time DNA evidence captured the public’s attention. But Strogatz, citing mathematician and author I.J. Good, focuses on a less well-known aspect of the case: whether O. J.’s history of abuse towards his wife was admissible into evidence.

The prosecution argued that abuse is often a precursor to murder, pointing to statistics indicating that an abuser was as much as ten times more likely to commit murder than was a random individual. The defense, however, countered with statistics indicating that the odds of an abusive husband murdering his wife were “infinitesimal,” roughly 1 in 2500. Based on those numbers, the actual relevance of a history of abuse to a murder case would appear limited at best. According to the defense, introducing that history would prejudice the jury against Simpson but would lack any probitive value, so the discussion should be barred.

In other words, both the defense and the prosecution were arguing conditional probability, specifically the likelihood that a woman will be murdered by her husband, given that her husband abuses her. But both defense and prosecution omitted a vital piece of data from their calculations: Nicole Brown Simpson *was* murdered. Strogatz points out that based on the defense’s numbers and the crime statistics of the time, the probability that a woman was murdered by her abuser, given that she was abused *and* murdered, is around 80%.

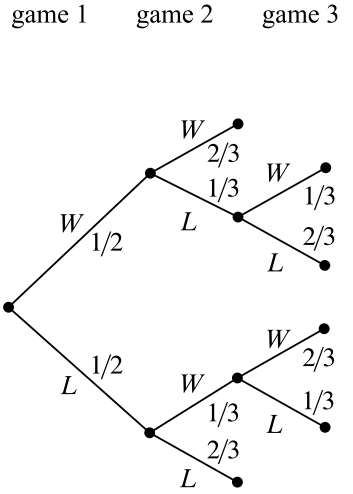
Strogatz’s article goes into more detail about the calculations behind that 80% figure. But the issue we want to illustrate is that conditional probability is used and misused all the time, and even experts under public scrutiny make mistakes.

18.3 The Four-Step Method for Conditional Probability

In a best-of-three tournament, the local C-league hockey team wins the first game with probability $1/2$. In subsequent games, their probability of winning is determined by the outcome of the previous game. If the local team won the previous game, then they are invigorated by victory and win the current game with probability $2/3$. If they lost the previous game, then they are demoralized by defeat and win the current game with probability only $1/3$. What is the probability that the local team wins the tournament, given that they win the first game?

This is a question about a conditional probability. Let A be the event that the local team wins the tournament, and let B be the event that they win the first game. Our goal is then to determine the conditional probability $\Pr[A \mid B]$.

We can tackle conditional probability questions just like ordinary probability problems: using a tree diagram and the four step method. A complete tree diagram is shown in Figure 18.1.



outcome	event A: win the series	event B: win game 1	outcome probability
WW	✓	✓	1/3
WLW	✓	✓	1/18
WLL		✓	1/9
LWW	✓		1/9
LWL			1/18
LL			1/3

Figure 18.1 The tree diagram for computing the probability that the local team wins two out of three games given that they won the first game.

Step 1: Find the Sample Space

Each internal vertex in the tree diagram has two children, one corresponding to a win for the local team (labeled W) and one corresponding to a loss (labeled L). The complete sample space is:

$$\mathcal{S} = \{WW, WLW, WLL, LWW, LWL, LL\}.$$

Step 2: Define Events of Interest

The event that the local team wins the whole tournament is:

$$T = \{WW, WLW, LWW\}.$$

And the event that the local team wins the first game is:

$$F = \{WW, WLW, WLL\}.$$

The outcomes in these events are indicated with check marks in the tree diagram in Figure 18.1.

Step 3: Determine Outcome Probabilities

Next, we must assign a probability to each outcome. We begin by labeling edges as specified in the problem statement. Specifically, the local team has a $1/2$ chance of winning the first game, so the two edges leaving the root are each assigned probability $1/2$. Other edges are labeled $1/3$ or $2/3$ based on the outcome of the preceding game. We then find the probability of each outcome by multiplying all probabilities along the corresponding root-to-leaf path. For example, the probability of outcome WLL is:

$$\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{9}.$$

Step 4: Compute Event Probabilities

We can now compute the probability that the local team wins the tournament, given that they win the first game:

$$\begin{aligned}
\Pr\left[A \mid B\right] &= \frac{\Pr[A \cap B]}{\Pr[B]} \\
&= \frac{\Pr[\{WW, WLW\}]}{\Pr[\{WW, WLW, WLL\}]} \\
&= \frac{1/3 + 1/18}{1/3 + 1/18 + 1/9} \\
&= \frac{7}{9}.
\end{aligned}$$

We're done! If the local team wins the first game, then they win the whole tournament with probability $7/9$.

18.4 Why Tree Diagrams Work

We've now settled into a routine of solving probability problems using tree diagrams, but we have not really explained why they work. The explanation is that the probabilities that we've been recording on the edges of tree diagrams are actually conditional probabilities.

For example, look at the uppermost path in the tree diagram for the hockey team problem, which corresponds to the outcome WW. The first edge is labeled $1/2$, which is the probability that the local team wins the first game. The second edge is labeled $2/3$, which is the probability that the local team wins the second game, *given* that they won the first—a conditional probability! More generally, on each edge of a tree diagram, we record the probability that the experiment proceeds along that path, given that it reaches the parent vertex.

So we've been using conditional probabilities all along. For example, we concluded that:

$$\Pr[WW] = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}.$$

Why is this correct?

The answer goes back to Definition 18.2.1 of conditional probability which could be written in a form called the *Product Rule* for conditional probabilities:

Rule (Conditional Probability Product Rule: 2 Events)**.**

$$\Pr[E_1 \cap E_2] = \Pr[E_1] \cdot \Pr\left[E_2 \mid E_1\right].$$

Multiplying edge probabilities in a tree diagram amounts to evaluating the right side of this equation. For example:

$$\begin{aligned}&\Pr[\text{win first game} \cap \text{win second game}]\\&\qquad = \Pr[\text{win first game}] \cdot \Pr\left[\text{win second game} \mid \text{win first game}\right]\\&\qquad = \frac{1}{2} \cdot \frac{2}{3}.\end{aligned}$$

So the Conditional Probability Product Rule is the formal justification for multiplying edge probabilities to get outcome probabilities.

To justify multiplying edge probabilities along a path of length three, we need a rule for three events:

Rule (Conditional Probability Product Rule: 3 Events)**.**

$$\Pr[E_1 \cap E_2 \cap E_3] = \Pr[E_1] \cdot \Pr\left[E_2 \mid E_1\right] \cdot \Pr\left[E_3 \mid E_1 \cap E_2\right].$$

An n-event version of the Rule is given in Problem 18.1, but its form should be clear from the three event version.

18.4.1 Probability of Size-k Subsets

As a simple application of the product rule for conditional probabilities, we can use the rule to calculate the number of size-k subsets of the integers $[1..n]$. Of course we already know this number is $\binom{n}{k}$, but now the rule will give us a new derivation of the formula for $\binom{n}{k}$.

Let’s pick some size-k subset $S \subseteq [1..n]$ as a target. Suppose we choose a size-k subset at random, with all subsets of $[1..n]$ equally likely to be chosen, and let p be the probability that our randomly chosen equals this target. That is, the probability of picking S is p, and since all sets are equally likely to be chosen, the number of size-k subsets equals $1/p$.

So what’s p? Well, the probability that the *smallest* number in the random set is one of the k numbers in S is k/n. Then, *given* that the smallest number in the random set is in S, the probability that the *second* smallest number in the random set is one of the remaining $k-1$ elements in S is $(k-1)/(n-1)$. So by the product rule, the probability that the *two* smallest numbers in the random set are both in S is

$$\frac{k}{n} \cdot \frac{k-1}{n-1}.$$

Next, given that the two smallest numbers in the random set are in S, the probability that the third smallest number is one of the $k-2$ remaining elements in S is $(k-2)/(n-2)$. So by the product rule, the probability that the *three* smallest numbers in the random set are all in S is

$$\frac{k}{n} \cdot \frac{k-1}{n-1} \cdot \frac{k-2}{n-2}.$$

Continuing in this way, it follows that the probability that *all* k elements in the randomly chosen set are in S, that is, the probabilty that the randomly chosen set equals the target, is

$$\begin{aligned}
p &= \frac{k}{n} \cdot \frac{k-1}{n-1} \cdot \frac{k-2}{n-2} \cdots \frac{k-(k-1)}{n-(k-1)} \\
&= \frac{k \cdot (k-1) \cdot (k-1) \cdots 1}{n \cdot (n-1) \cdot (n-2) \cdots (n-(k-1))} \\
&= \frac{k!}{n!/(n-k)!} \\
&= \frac{k!(n-k)!}{n!}.
\end{aligned}$$

So we have again shown the number of size-k subsets of $[1..n]$, namely $1/p$, is

$$\frac{n!}{k!(n-k)!}.$$

18.4.2 Medical Testing

Breast cancer is a deadly disease that claims thousands of lives every year. Early detection and accurate diagnosis are high priorities, and routine mammograms are one of the first lines of defense. They're not very accurate as far as medical tests go, but they are correct between 90% and 95% of the time, which seems pretty good for a relatively inexpensive non-invasive test.[1] However, mammogram results are also an example of conditional probabilities having counterintuitive consequences. If the test was positive for breast cancer in you or a loved one, and the test is better than 90% accurate, you'd naturally expect that to mean there is better than 90% chance that the disease was present. But a mathematical analysis belies that naive intuitive expectation. Let's start by precisely defining how accurate a mammogram is:

- If you have the condition, there is a 10% chance that the test will say you do not have it. This is called a "false negative."
- If you do not have the condition, there is a 5% chance that the test will say you do. This is a "false positive."

18.4.3 Four Steps Again

Now suppose that we are testing middle-aged women with no family history of cancer. Among this cohort, incidence of breast cancer rounds up to about 1%.

Step 2: Define Events of Interest
Let A be the event that the person has breast cancer. Let B be the event that the test was positive. The outcomes in each event are marked in the tree diagram. We want to find $\Pr\left[A \mid B\right]$, the probability that a person has breast cancer, given that the test was positive.

Step 3: Find Outcome Probabilities
First, we assign probabilities to edges. These probabilities are drawn directly from the problem statement. By the Product Rule, the probability of an outcome is the product of the probabilities on the corresponding root-to-leaf path. All probabilities are shown in Figure 18.2.

Step 4: Compute Event Probabilities
From Definition 18.2.1, we have

$$\Pr\left[A \mid B\right] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{0.009}{0.009 + 0.0495} \approx 15.4\%.$$

So, if the test is positive, then there is an 84.6% chance that the result is incorrect, even though the test is nearly 95% accurate! So this seemingly pretty accurate

[1] The statistics in this example are roughly based on actual medical data, but have been altered somewhat to simplify the calculations.

Step 1: Find the Sample Space

The sample space is found with the tree diagram in Figure 18.2.

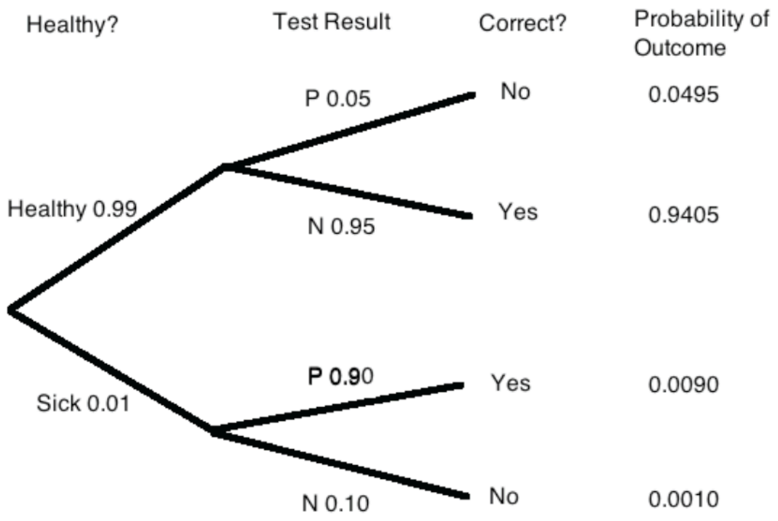


Figure 18.2 The tree diagram for a breast cancer test.

test doesn’t tell us much. To see why percent accuracy is no guarantee of value, notice that there is a simple way to make a test that is 99% accurate: always return a negative result! This test gives the right answer for all healthy people and the wrong answer only for the 1% that actually have cancer. This 99% accurate test tells us nothing; the “less accurate” mammogram is still a lot more useful.

18.4.4 Natural Frequencies

That there is only about a 15% chance that the patient actually has the condition when the test say so may seem surprising at first, but it makes sense with a little thought. There are two ways the patient could test positive: first, the patient could have the condition and the test could be correct; second, the patient could be healthy and the test incorrect. But almost everyone is healthy! The number of healthy individuals is so large that even the mere 5% with false positive results overwhelm the number of genuinely positive results from the truly ill.

Thinking like this in terms of these “natural frequencies” can be a useful tool for interpreting some of the strange seeming results coming from those formulas. For example, let’s take a closer look at the mammogram example.

Imagine 10,000 women in our demographic. Based on the frequency of the disease, we’d expect 100 of them to have breast cancer. Of those, 90 would have a positive result. The remaining 9,900 woman are healthy, but 5% of them—500, give or take—will show a false positive on the mammogram. That gives us 90 real positives out of a little fewer than 600 positives. An 85% error rate isn’t so surprising after all.

18.4.5 *A Posteriori* Probabilities

If you think about it much, the medical testing problem we just considered could start to trouble you. You may wonder if a statement like “If someone tested positive, then that person has the condition with probability 18%” makes sense, since a given person being tested either has the disease or they don’t.

One way to understand such a statement is that it just means that 15% of the people who test positive will actually have the condition. Any particular person has it or they don’t, but a *randomly selected* person among those who test positive will have the condition with probability 15%.

But what does this 15% probability tell you if you *personally* got a positive result? Should you be relieved that there is less than one chance in five that you have the disease? Should you worry that there is nearly one chance in five that you do have the disease? Should you start treatment just in case? Should you get more tests?

These are crucial practical questions, but it is important to understand that they are not *mathematical* questions. Rather, these are questions about statistical judgements and the philosophical meaning of probability. We'll say a bit more about this after looking at one more example of after-the-fact probabilities.

The Hockey Team in Reverse

Suppose that we turn the hockey question around: what is the probability that the local C-league hockey team won their first game, given that they won the series?

As we discussed earlier, some people find this question absurd. If the team has already won the tournament, then the first game is long since over. Who won the first game is a question of fact, not of probability. However, our mathematical theory of probability contains no notion of one event preceding another. There is no notion of time at all. Therefore, from a mathematical perspective, this is a perfectly valid question. And this is also a meaningful question from a practical perspective. Suppose that you're told that the local team won the series, but not told the results of individual games. Then, from your perspective, it makes perfect sense to wonder how likely it is that local team won the first game.

A conditional probability $\Pr\left[B \mid A\right]$ is called *a posteriori* if event B precedes event A in time. Here are some other examples of a posteriori probabilities:

- The probability it was cloudy this morning, given that it rained in the afternoon.

- The probability that I was initially dealt two queens in Texas No Limit Hold 'Em poker, given that I eventually got four-of-a-kind.

from ordinary probabilities; the distinction comes from our view of causality, which is a philosophical question rather than a mathematical one.

Let's return to the original problem. The probability that the local team won their first game, given that they won the series is $\Pr\left[B \mid A\right]$. We can compute this using the definition of conditional probability and the tree diagram in Figure 18.1:

$$\Pr\left[B \mid A\right] = \frac{\Pr[B \cap A]}{\Pr[A]} = \frac{1/3 + 1/18}{1/3 + 1/18 + 1/9} = \frac{7}{9}.$$

In general, such pairs of probabilities are related by Bayes' Rule:

Theorem 18.4.1 (Bayes' Rule)**.**

$$\Pr\left[B \mid A\right] = \frac{\Pr\left[A \mid B\right] \cdot \Pr[B]}{\Pr[A]} \tag{18.2}$$

Proof. We have

$$\Pr\left[B \mid A\right] \cdot \Pr[A] = \Pr[A \cap B] = \Pr\left[A \mid B\right] \cdot \Pr[B]$$

by definition of conditional probability. Dividing by $\Pr[A]$ gives (18.2). ■

18.4.6 Philosphy of Probability

Let's try to assign a probability to the event

$$[2^{6972607} - 1 \text{ is a prime number}]$$

It's not obvious how to check whether such a large number is prime, so you might try an estimation based on the density of primes. The Prime Number Theorem implies that only about 1 in 5 million numbers in this range are prime, so you might say that the probability is about $2 \cdot 10^{-8}$. On the other hand, given that we chose this example to make some philosophical point, you might guess that we probably purposely chose an obscure looking prime number, and you might be willing to make an even money bet that the number is prime. In other words, you might think the probability is 1/2. Finally, we can take the position that assigning a probability to this statement is nonsense because there is no randomness involved; the number is either prime or it isn't. This is the view we take in this text.

An alternate view is the *Bayesian* approach, in which a probability is interpreted as a *degree of belief* in a proposition. A Bayesian would agree that the number above is either prime or composite, but they would be perfectly willing to assign a probability to each possibility. The Bayesian approach is very broad in its willingness to assign probabilities to any event, but the problem is that there is no single "right" probability for an event, since the probability depends on one's initial beliefs. On the other hand, if you have confidence in some set of initial beliefs, then Bayesianism provides a convincing framework for updating your beliefs as further information emerges.

As an aside, it is not clear whether Bayes himself was Bayesian in this sense. However, a Bayesian would be willing to talk about the probability that Bayes was Bayesian.

Another school of thought says that probabilities can only be meaningfully applied to *repeatable processes* like rolling dice or flipping coins. In this *frequentist* view, the probability of an event represents the fraction of trials in which the event occurred. So we can make sense of the *a posteriori* probabilities of the C-league hockey example of Section 18.4.5 by imagining that many hockey series were played, and the probability that the local team won their first game, given that they won the series, is simply the fraction of series where they won the first game among all the series they won.

Getting back to prime numbers, we mentioned in Section 9.5.1 that there is a probabilistic primality test. If a number N is composite, there is at least a $3/4$ chance that the test will discover this. In the remaining $1/4$ of the time, the test is inconclusive. But as long as the result is inconclusive, the test can be run independently again and again up to, say, 100 times. So if N actually is composite, then the probability that 100 repetitions of the probabilistic test do not discover this is at most:

$$\left(\frac{1}{4}\right)^{100}.$$

If the test remained inconclusive after 100 repetitions, it is still logically possible that N is composite, but betting that N is prime would be the best bet you'll ever get to make! If you're comfortable using probability to describe your personal belief about primality after such an experiment, you are being a Bayesian. A frequentist would not assign a probability to N's primality, but they would also be happy to bet on primality with tremendous *confidence*. We'll examine this issue again when we discuss polling and confidence levels in Section 18.9.

Despite the philosophical divide, the real world conclusions Bayesians and Frequentists reach from probabilities are pretty much the same, and even where their interpretations differ, they use the same theory of probability.

18.5 The Law of Total Probability

Breaking a probability calculation into cases simplifies many problems. The idea is to calculate the probability of an event A by splitting into two cases based on whether or not another event E occurs. That is, calculate the probability of $A \cap E$ and $A \cap \overline{E}$. By the Sum Rule, the sum of these probabilities equals $\Pr[A]$. Expressing the intersection probabilities as conditional probabilities yields:

Rule 18.5.1 (Law of Total Probability: single event)**.**

$$\Pr[A] = \Pr\left[A \mid E\right] \cdot \Pr[E] + \Pr\left[A \;\middle|\; \overline{E}\right] \cdot \Pr[\overline{E}].$$

For example, suppose we conduct the following experiment. First, we flip a fair coin. If heads comes up, then we roll one die and take the result. If tails comes up, then we roll two dice and take the sum of the two results. What is the probability that this process yields a 2? Let E be the event that the coin comes up heads, and let A be the event that we get a 2 overall. Assuming that the coin is fair, $\Pr[E] = \Pr[\overline{E}] = 1/2$. There are now two cases. If we flip heads, then we roll a 2 on a single die with probability $\Pr\left[A \mid E\right] = 1/6$. On the other hand, if we flip tails, then we get a sum of 2 on two dice with probability $\Pr\left[A \;\middle|\; \overline{E}\right] = 1/36$. Therefore, the probability that the whole process yields a 2 is

$$\Pr[A] = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{36} = \frac{7}{72}.$$

This rule extends to any set of disjoint events that make up the entire sample space. For example,

Rule (Law of Total Probability: 3-events)**.** *If* E_1, E_2 *and* E_3 *are disjoint, and* $\Pr[E_1 \cup E_2 \cup E_3] = 1$*, then*

$$\Pr[A] = \Pr\left[A \mid E_1\right] \cdot \Pr[E_1] + \Pr\left[A \mid E_2\right] \cdot \Pr[E_2] + \Pr\left[A \mid E_3\right] \cdot \Pr[E_3].$$

This in turn leads to a three-event version of Bayes' Rule in which the probability of event E_1 given A is calculated from the "inverse" conditional probabilities of A given E_1, E_2, and E_3:

Rule (Bayes' Rule: 3-events)**.**

$$\Pr\left[E_1 \mid A\right] = \frac{\Pr\left[A \mid E_1\right] \cdot \Pr[E_1]}{\Pr\left[A \mid E_1\right] \cdot \Pr[E_1] + \Pr\left[A \mid E_2\right] \cdot \Pr[E_2] + \Pr\left[A \mid E_3\right] \cdot \Pr[E_3]}$$

The generalization of these rules to n disjoint events is a routine exercise (Problems 18.3 and 18.4).

18.5.1 Conditioning on a Single Event

The probability rules that we derived in Section 17.5.2 extend to probabilities conditioned on the same event. For example, the Inclusion-Exclusion formula for two sets holds when all probabilities are conditioned on an event C:

$$\Pr\left[A \cup B \mid C\right] = \Pr\left[A \mid C\right] + \Pr\left[B \mid C\right] - \Pr\left[A \cap B \mid C\right].$$

This is easy to verify by plugging in the Definition 18.2.1 of conditional probability.[2]

It is important not to mix up events before and after the conditioning bar. For example, the following is *not* a valid identity:

False Claim.

$$\Pr\left[A \mid B \cup C\right] = \Pr\left[A \mid B\right] + \Pr\left[A \mid C\right] - \Pr\left[A \mid B \cap C\right]. \qquad (18.3)$$

A simple counter-example is to let B and C be events over a uniform space with most of their outcomes in A, but not overlapping. This ensures that $\Pr\left[A \mid B\right]$ and $\Pr\left[A \mid C\right]$ are both close to 1. For example,

$$\begin{aligned} B &::= [0..9], \\ C &::= [10..18] \cup \{0\}, \\ A &::= [1..18], \end{aligned}$$

[2] Problem 18.14 explains why this and similar conditional identities follow on general principles from the corresponding unconditional identities.

so

$$\Pr\left[A \mid B\right] = \frac{9}{10} = \Pr\left[A \mid C\right].$$

Also, since 0 is the only outcome in $B \cap C$ and $0 \notin A$, we have

$$\Pr\left[A \mid B \cap C\right] = 0$$

So the right-hand side of (18.3) is 1.8, while the left-hand side is a probability which can be at most 1—actually, it is 18/19.

18.6 Simpson's Paradox

In 1973, a famous university was investigated for gender discrimination [7]. The investigation was prompted by evidence that, at first glance, appeared definitive: in 1973, 44% of male applicants to the school's graduate programs were accepted, but only 35% of female applicants were admitted.

However, this data turned out to be completely misleading. Analysis of the individual departments, showed not only that few showed significant evidence of bias, but also that among the few departments that *did* show statistical irregularities, most were slanted *in favor of women*. This suggests that if there was any sex discrimination, then it was against men!

Given the discrepancy in these findings, it feels like someone must be doing bad math—intentionally or otherwise. But the numbers are not actually inconsistent. In fact, this statistical hiccup is common enough to merit its own name: *Simpson's Paradox* occurs when multiple small groups of data all exhibit a similar trend, but that trend reverses when those groups are aggregated. To explain how this is possible, let's first clarify the problem by expressing both arguments in terms of conditional probabilities. For simplicity, suppose that there are only two departments EE and CS. Consider the experiment where we pick a random candidate. Define the following events:

- A ::= the candidate is admitted to his or her program of choice,
- F_{EE} ::= the candidate is a woman applying to the EE department,
- F_{CS} ::= the candidate is a woman applying to the CS department,
- M_{EE} ::= the candidate is a man applying to the EE department,
- M_{CS} ::= the candidate is a man applying to the CS department.

CS	2 men admitted out of 5 candidates	40%
	50 women admitted out of 100 candidates	50%
EE	70 men admitted out of 100 candidates	70%
	4 women admitted out of 5 candidates	80%
Overall	72 men admitted, 105 candidates	$\approx$ 69%
	54 women admitted, 105 candidates	$\approx$ 51%

Table 18.1 Hypothetical admission statistics where men are overall more to be admitted, but are less likely to be admitted into each department.

Assume that all candidates are either men or women, and that no candidate belongs to both departments. That is, the events F_{EE}, F_{CS}, M_{EE} and M_{CS} are all disjoint.

In these terms, the plaintiff's assertion—that a male candidate is more likely to be admitted to the university than a female—can be expressed by the following inequality:

$$\Pr[A \mid M_{EE} \cup M_{CS}] > \Pr[A \mid F_{EE} \cup F_{CS}].$$

The university's retort that *in any given department*, a male applicant is less likely to be admitted than a female can be expressed by a pair of inequalities:

$$\Pr[A \mid M_{EE}] < \Pr[A \mid F_{EE}] \quad \text{and}$$
$$\Pr[A \mid M_{CS}] < \Pr[A \mid F_{CS}].$$

We can explain how there could be such a discrepancy between university-wide and department-by-department admission statistics by supposing that the CS department is more selective than the EE department, but CS attracts a far larger number of woman applicants than EE.[3]. Table 18.1 shows some admission statistics contrived to highlight how the inequalities asserted by both the plaintiff and the university could both hold.

Initially, we and the plaintiffs both assumed that the overall admissions statistics for the university could only be explained by gender discrimination. The department by department statistics seems to belie the accusation of discrimination. But do they really?

Suppose we replaced "the candidate is a man/woman applying to the EE department," by "the candidate is a man/woman for whom an admissions decision was made during an odd-numbered day of the month," and likewise with CS and an even-numbered day of the month. Since we don't think the parity of a date is a

[3] At the actual university in the lawsuit, the "exclusive" departments more popular among women were those that did not require a mathematical foundation, such as English and education. Women's disproportionate choice of these careers reflects gender bias, but one which predates the university's involvement.

cause for the outcome of an admission decision, we would most likely dismiss the "coincidence" that on both odd and even dates, women are more frequently admitted. Instead we would judge, based on the overall data showing women less likely to be admitted, that gender bias against women *was* an issue in the university.

Bear in mind that it would be the *same numerical data* that we would be using to justify our different conclusions in the department-by-department case and the even-day-odd-day case. We interpreted the same numbers differently based on our implicit causal beliefs, specifically that departments matter and date parity does not. It is circular to claim that the data corroborated our beliefs that there is or is not discrimination. Rather, our interpretation of the data correlation depended on our beliefs about the causes of admission in the first place.[4] This example highlights a basic principle in statistics that people constantly ignore: *never assume that correlation implies causation.*

18.7 Independence

Suppose that we flip two fair coins simultaneously on opposite sides of a room. Intuitively, the way one coin lands does not affect the way the other coin lands. The mathematical concept that captures this intuition is called *independence*.

Definition 18.7.1. An event with probability 0 is defined to be independent of every event (including itself). If $\Pr[B] \neq 0$, then event A is independent of event B iff

$$\Pr\left[A \mid B\right] = \Pr[A]. \tag{18.4}$$

In other words, A and B are independent if knowing that B happens does not alter the probability that A happens, as is the case with flipping two coins on opposite sides of a room.

Potential Pitfall

Students sometimes get the idea that disjoint events are independent. The *opposite* is true: if $A \cap B = \emptyset$, then knowing that A happens means you know that B does not happen. Disjoint events are *never* independent—unless one of them has probability zero.

[4] These issues are thoughtfully examined in *Causality: Models, Reasoning and Inference*, Judea Pearl, Cambridge U. Press, 2001.

18.7.1 Alternative Formulation

Sometimes it is useful to express independence in an alternate form which follows immediately from Definition 18.7.1:

Theorem 18.7.2. *A is independent of B if and only if*

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]. \tag{18.5}$$

Notice that Theorem 18.7.2 makes apparent the symmetry between A being independent of B and B being independent of A:

Corollary 18.7.3. *A is independent of B iff B is independent of A.*

18.7.2 Independence Is an Assumption

Generally, independence is something that you *assume* in modeling a phenomenon. For example, consider the experiment of flipping two fair coins. Let A be the event that the first coin comes up heads, and let B be the event that the second coin is heads. If we assume that A and B are independent, then the probability that both coins come up heads is:

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

In this example, the assumption of independence is reasonable. The result of one coin toss should have negligible impact on the outcome of the other coin toss. And if we were to repeat the experiment many times, we would be likely to have $A \cap B$ about 1/4 of the time.

On the other hand, there are many examples of events where assuming independence isn’t justified. For example, an hourly weather forecast for a clear day might list a 10% chance of rain every hour from noon to midnight, meaning each hour has a 90% chance of being dry. But that does *not* imply that the odds of a rainless day are a mere $0.9^{12} \approx 0.28$. In reality, if it doesn’t rain as of 5pm, the odds are higher than 90% that it will stay dry at 6pm as well—and if it starts pouring at 5pm, the chances are much higher than 10% that it will still be rainy an hour later.

Deciding when to *assume* that events are independent is a tricky business. In practice, there are strong motivations to assume independence since many useful formulas—such as equation (18.5)—only hold if the events are independent. But you need to be careful: we’ll describe several famous examples where mistaken assumptions of independence led to trouble. This problem gets even trickier when there are more than two events in play.

18.8 Mutual Independence

We have defined what it means for two events to be independent. What if there are more than two events? For example, how can we say that the flips of n coins are all independent of one another? A set of events is said to be *mutually independent* if the probability of each event in the set is the same no matter which of the other events has occurred. This is equivalent to saying that for any selection of two or more of the events, the probability that all the selected events occur equals the product of the probabilities of the selected events.

For example, four events E_1, E_2, E_3, E_4 are mutually independent if and only if all of the following equations hold:

$$\begin{aligned}
\Pr[E_1 \cap E_2] &= \Pr[E_1] \cdot \Pr[E_2] \\
\Pr[E_1 \cap E_3] &= \Pr[E_1] \cdot \Pr[E_3] \\
\Pr[E_1 \cap E_4] &= \Pr[E_1] \cdot \Pr[E_4] \\
\Pr[E_2 \cap E_3] &= \Pr[E_2] \cdot \Pr[E_3] \\
\Pr[E_2 \cap E_4] &= \Pr[E_2] \cdot \Pr[E_4] \\
\Pr[E_3 \cap E_4] &= \Pr[E_3] \cdot \Pr[E_4] \\
\Pr[E_1 \cap E_2 \cap E_3] &= \Pr[E_1] \cdot \Pr[E_2] \cdot \Pr[E_3] \\
\Pr[E_1 \cap E_2 \cap E_4] &= \Pr[E_1] \cdot \Pr[E_2] \cdot \Pr[E_4] \\
\Pr[E_1 \cap E_3 \cap E_4] &= \Pr[E_1] \cdot \Pr[E_3] \cdot \Pr[E_4] \\
\Pr[E_2 \cap E_3 \cap E_4] &= \Pr[E_2] \cdot \Pr[E_3] \cdot \Pr[E_4] \\
\Pr[E_1 \cap E_2 \cap E_3 \cap E_4] &= \Pr[E_1] \cdot \Pr[E_2] \cdot \Pr[E_3] \cdot \Pr[E_4]
\end{aligned}$$

The generalization to mutual independence of n events should now be clear.

18.8.1 DNA Testing

Assumptions about independence are routinely made in practice. Frequently, such assumptions are quite reasonable. Sometimes, however, the reasonableness of an independence assumption is not so clear, and the consequences of a faulty assumption can be severe.

Let’s return to the O. J. Simpson murder trial. The following expert testimony was given on May 15, 1995:

Mr. Clarke: When you make these estimations of frequency—and I believe you touched a little bit on a concept called independence?

Dr. Cotton: Yes, I did.

Mr. Clarke: And what is that again?

Dr. Cotton: It means whether or not you inherit one allele that you have is not—does not affect the second allele that you might get. That is, if you inherit a band at 5,000 base pairs, that doesn't mean you'll automatically or with some probability inherit one at 6,000. What you inherit from one parent is [independent of] what you inherit from the other.

Mr. Clarke: Why is that important?

Dr. Cotton: Mathematically that's important because if that were not the case, it would be improper to multiply the frequencies between the different genetic locations.

Mr. Clarke: How do you—well, first of all, are these markers independent that you've described in your testing in this case?

Presumably, this dialogue was as confusing to you as it was for the jury. Essentially, the jury was told that genetic markers in blood found at the crime scene matched Simpson's. Furthermore, they were told that the probability that the markers would be found in a randomly-selected person was at most 1 in 170 million. This astronomical figure was derived from statistics such as:

- 1 person in 100 has marker A.
- 1 person in 50 marker B.
- 1 person in 40 has marker C.
- 1 person in 5 has marker D.
- 1 person in 170 has marker E.

Then these numbers were multiplied to give the probability that a randomly-selected person would have all five markers:

$$\begin{aligned}\Pr[A \cap B \cap C \cap D \cap E] &= \Pr[A] \cdot \Pr[B] \cdot \Pr[C] \cdot \Pr[D] \cdot \Pr[E] \\ &= \frac{1}{100} \cdot \frac{1}{50} \cdot \frac{1}{40} \cdot \frac{1}{5} \cdot \frac{1}{170} = \frac{1}{170{,}000{,}000}.\end{aligned}$$

The defense pointed out that this assumes that the markers appear mutually independently. Furthermore, all the statistics were based on just a few hundred blood samples.

After the trial, the jury was widely mocked for failing to "understand" the DNA evidence. If you were a juror, would *you* accept the 1 in 170 million calculation?

18.8.2 Pairwise Independence

The definition of mutual independence seems awfully complicated—there are so many selections of events to consider! Here's an example that illustrates the subtlety of independence when more than two events are involved. Suppose that we flip three fair, mutually-independent coins. Define the following events:

- A_1 is the event that coin 1 matches coin 2.
- A_2 is the event that coin 2 matches coin 3.
- A_3 is the event that coin 3 matches coin 1.

Are A_1, A_2, A_3 mutually independent?

The sample space for this experiment is:

$$\{HHH,\ HHT,\ HTH,\ HTT,\ THH,\ THT,\ TTH,\ TTT\}.$$

Every outcome has probability $(1/2)^3 = 1/8$ by our assumption that the coins are mutually independent.

To see if events A_1, A_2 and A_3 are mutually independent, we must check a sequence of equalities. It will be helpful first to compute the probability of each event A_i:

$$\begin{aligned}\Pr[A_1] &= \Pr[HHH] + \Pr[HHT] + \Pr[TTH] + \Pr[TTT]\\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}.\end{aligned}$$

By symmetry, $\Pr[A_2] = \Pr[A_3] = 1/2$ as well. Now we can begin checking all the equalities required for mutual independence:

$$\begin{aligned}\Pr[A_1 \cap A_2] &= \Pr[HHH] + \Pr[TTT] = \frac{1}{8} + \frac{1}{8} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}\\ &= \Pr[A_1]\Pr[A_2].\end{aligned}$$

By symmetry, $\Pr[A_1 \cap A_3] = \Pr[A_1] \cdot \Pr[A_3]$ and $\Pr[A_2 \cap A_3] = \Pr[A_2] \cdot \Pr[A_3]$ must hold also. Finally, we must check one last condition:

$$\begin{aligned}\Pr[A_1 \cap A_2 \cap A_3] &= \Pr[HHH] + \Pr[TTT] = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}\\ &\neq \frac{1}{8} = \Pr[A_1]\Pr[A_2]\Pr[A_3].\end{aligned}$$

The three events A_1, A_2 and A_3 are not mutually independent even though any two of them are independent! This not-quite mutual independence seems weird at first, but it happens. It even generalizes:

Definition 18.8.1. A set $A_1, A_2, \ldots,$ of events is *k-way independent* iff every set of k of these events is mutually independent. The set is *pairwise independent* iff it is 2-way independent.

So the events A_1, A_2, A_3 above are pairwise independent, but not mutually independent. Pairwise independence is a much weaker property than mutual independence.

For example, suppose that the prosecutors in the O. J. Simpson trial were wrong and markers A, B, C, D and E are only *pairwise* independently. Then the probability that a randomly-selected person has all five markers is no more than:

$$\begin{aligned}\Pr[A \cap B \cap C \cap D \cap E] &\le \Pr[A \cap E] = \Pr[A] \cdot \Pr[E] \\ &= \frac{1}{100} \cdot \frac{1}{170} = \frac{1}{17{,}000}.\end{aligned}$$

The first line uses the fact that $A \cap B \cap C \cap D \cap E$ is a subset of $A \cap E$. (We picked out the A and E markers because they're the rarest.) We use pairwise independence on the second line. Now the probability of a random match is 1 in 17,000—a far cry from 1 in 170 million! And this is the strongest conclusion we can reach assuming only pairwise independence.

On the other hand, the 1 in 17,000 bound that we get by assuming pairwise independence is a lot better than the bound that we would have if there were no independence at all. For example, if the markers are dependent, then it is possible that

everyone with marker E has marker A,

everyone with marker A has marker B,

everyone with marker B has marker C, and

everyone with marker C has marker D.

In such a scenario, the probability of a match is

$$\Pr[E] = \frac{1}{170}.$$

So a stronger independence assumption leads to a smaller bound on the probability of a match. The trick is to figure out what independence assumption is reasonable. Assuming that the markers are *mutually* independent may well *not* be reasonable unless you have examined hundreds of millions of blood samples. Otherwise, how would you know that marker D does not show up more frequently whenever the other four markers are simultaneously present?

18.9 Probability versus Confidence

Let's look at some other problems like the breast cancer test of Section 18.4.2, but this time we'll use more extreme numbers to highlight some key issues.

18.9.1 Testing for Tuberculosis

Let's suppose we have a really terrific diagnostic test for tuberculosis (TB): if you have TB, the test is *guaranteed* to detect it, and if you don't have TB, then the test will report that correctly 99% of the time!

In other words, let "TB" be the event that a person has TB, "pos" be the event that the person tests positive for TB, so "$\overline{pos}$" is the event that they test negative. Now we can restate these guarantees in terms of conditional probabilities:

$$\Pr\left[pos \mid TB\right] = 1, \tag{18.6}$$

$$\Pr\left[\overline{pos} \mid \overline{TB}\right] = 0.99. \tag{18.7}$$

This means that the test produces the correct result at least 99% of the time, regardless of whether or not the person has TB. A careful statistician would assert:[5]

Lemma. *You can be 99% confident that the test result is correct.*

Corollary 18.9.1. *If you test positive, then*

> ***either** you have TB **or** something very unlikely (probability 1/100) happened.*

Lemma 18.9.1 and Corollary 18.9.1 may *seem* to be saying that

False Claim. *If you test positive, then the probability that you have TB is* 0.99*.*

But this would be a mistake.

To highlight the difference between confidence in the test diagnosis versus the probability of TB, let's think about what to do if you test positive. Corollary 18.9.1

[5] Confidence is usually used to describe the probability that a statistical estimations of some quantity is correct (Section 20.4.3). We are trying to simplify the discussion by using this one concept to illustrate standard approaches to both hypothesis testing and estimation.

In the context of hypothesis testing, statisticians would normally distinguish the "false positive" probability, in this case the probability 0.01 that a healthy person is incorrectly diagnosed as having TB, and call this the *significance* of the test. The "false negative" probability would be the probability that person with TB is incorrectly diagnosed as healthy; it is zero. The *power* of the test is one minus the false negative probability, so in this case the power is the highest possible, namely, one.

seems to suggest that it's worth betting with high odds that you have TB, because it makes sense to bet against something unlikely happening—like the test being wrong. But having TB actually turns out to be *a lot less likely* than the test being wrong. So the either-or of Corollary 18.9.1 is really an either-or between something happening that is *extremely* unlikely—having TB—and something that is only *very* unlikely—the diagnosis being wrong. You're better off betting against the extremely unlikely event: it is better to bet the diagnosis is wrong.

So some knowledge of the probability of having TB is needed in order to figure out how seriously to take a positive diagnosis, even when the diagnosis is given with what seems like a high level of confidence. We can see exactly how the frequency of TB in a population influences the importance of a positive diagnosis by actually calculating the probability that someone who tests positive has TB. That is, we want to calculate $\Pr\left[TB \mid pos\right]$, which we do next.

18.9.2 Updating the Odds

Bayesian Updating

A standard way to convert the test probabilities into outcome probabilities is to use Bayes Theorem (18.2). It will be helpful to rephrase Bayes Theorem in terms of "odds" instead of probabilities.

If H is an event, we define the *odds* of H to be

$$\text{Odds}(H) ::= \frac{\Pr[H]}{\Pr[\overline{H}]} = \frac{\Pr[H]}{1 - \Pr[H]}.$$

For example, if H is the event of rolling a four using a fair, six-sided die, then

$$\Pr[\text{roll four}] = 1/6, \text{ so}$$
$$\text{Odds}(\text{roll four}) = \frac{1/6}{5/6} = \frac{1}{5}.$$

A gambler would say the odds of rolling a four were "one to five," or equivalently, "five to one *against*" rolling a four.

Odds are just another way to talk about probabilities. For example, saying the odds that a horse will win a race are "three to one against" means that a winning \$1 bet will return \$3 plus the \$1 bet initially. Three to one against winnning is the same as odds of one to three that the horse will win, which means the horse will win with probability $1/4$. In general,

$$\Pr[H] = \frac{\text{Odds}(H)}{1 + \text{Odds}(H)}.$$

Now suppose an event E offers some evidence about H. We now want to find the conditional probability of H given E. We can just as well find the odds of H given E,

$$\begin{aligned}
\text{Odds}(H \mid E) &::= \frac{\Pr\left[H \mid E\right]}{\Pr\left[\overline{H} \mid E\right]} \\
&= \frac{\Pr\left[E \mid H\right]\Pr[H]/\Pr[E]}{\Pr\left[E \mid \overline{H}\right]\Pr[\overline{H}]/\Pr[E]} && \text{(Bayes Theorem)} \\
&= \frac{\Pr\left[E \mid H\right]}{\Pr\left[E \mid \overline{H}\right]} \cdot \frac{\Pr[H]}{\Pr[\overline{H}]} \\
&= \text{Bayes-factor}(E, H) \cdot \text{Odds}(H),
\end{aligned}$$

where

$$\text{Bayes-factor}(E, H) ::= \frac{\Pr\left[E \mid H\right]}{\Pr\left[E \mid \overline{H}\right]}.$$

So to update the odds of H given the evidence E, we just multiply by Bayes Factor:

Lemma 18.9.2.

$$\text{Odds}(H \mid E) = \text{Bayes-factor}(E, H) \cdot \text{Odds}(H).$$

Odds for the TB test

The probabilities of test outcomes given in (18.6) and (18.7) are exactly what we need to find Bayes factor for the TB test:

$$\begin{aligned}
\text{Bayes-factor}(pos, TB) &= \frac{\Pr\left[pos \mid TB\right]}{\Pr\left[pos \mid \overline{TB}\right]} \\
&= \frac{1}{1 - \Pr\left[\overline{pos} \mid \overline{TB}\right]} \\
&= \frac{1}{1 - 0.99} = 100.
\end{aligned}$$

So testing positive for TB increases the odds you have TB by a factor of 100, which means a positive test is significant evidence supporting a diagnosis of TB. That seems good to know. But Lemma 18.9.2 also makes it clear that when a random person tests positive, we still can't determine the odds they have TB unless we know what are the *odds of their having TB in the first place*, so let's examine that.

In 2011, the United States Center for Disease Control got reports of 11,000 cases of TB in US. We can estimate that there were actually about 30,000 cases of TB that year, since it seems that only about one third of actual cases of TB get reported. The US population is a little over 300 million, which means

$$\Pr[TB] \approx \frac{30,000}{300,000,000} = \frac{1}{10,000}.$$

So the odds of TB are $1/9999$. Therefore,

$$\text{Odds}(TB \mid pos) = 100 \cdot \frac{1}{9,999} \approx \frac{1}{100}.$$

In other words, even if someone tests positive for TB at the 99% confidence level, the odds remain about 100 to one *against* their having TB. The 99% confidence level is not nearly high enough to overcome the relatively tiny probability of having TB.

18.9.3 Facts that are Probably True

We have figured out that if a random person tests positive for TB, the probability they have TB is about 1/100. Now if you personally happened to test positive for TB, a competent doctor typically would tell you that the probability that you have TB has risen from 1/10,000 to 1/100. But has it? Not really.

Your doctor should have not have been talking in this way about your particular situation. He should just have stuck to the statement that for *randomly chosen* people, the positive test would be right only one percent of the time. But you are not a random person, and whether or not you have TB is a fact about reality. The truth about your having TB may be *unknown* to your doctor and you, but that does not mean it has some probability of being true. It is either true or false, we just don’t know which.

In fact, if you were worried about a 1/100 probability of having this serious disease, you could use additional information about yourself to change this probability. For example, native born residents of the US are about half as likely to have TB as foreign born residents. So if you are native born, “your” probability of having TB halves. Conversely, TB is twenty-five times more frequent among native born Asian/Pacific Islanders than native born Caucasions. So your probability of TB would increase dramatically if your family was from an Asian/Pacific Island.

The point is that the probability of having TB that your doctor reports to you depends on the probability of TB for a random person whom the doctor thinks is *like you*. The doctor has made a judgment about you based, for example, on what personal factors he considers relevant to getting TB, or how serious he thinks the consequences of a mistaken diagnosis would be. These are important medical judgments, but they are not mathematical. Different doctors will make different judgments about who is like you, and they will report differing probabilities. There is no "true" model of who you are, and there is no true individual probability of your having TB.

18.9.4 Extreme events

By definition, flipping a *fair* coin is equally likely to come up Heads or Tails. Now suppose you flip a fair coin one hundred times and get a Head every time. What do you think the odds are that the next flip will also be a Head?

If we make the usual assumption that the coin remains fair after one hundred flips, then by definition the official answer is that a Tail on the next flip is just as likely as another Head. But this belies what any sensible person would do, which is to bet heavily on the next flip being another Head.

How to make sense of this? To begin, let's recognize how absurd it is to wonder about what happens after one hundred flips of a fair coin all come up Heads, because the probability of this happening is unimaginably tiny. For example, the probability that just *fifty* flips of a fair coin come up Heads is 2^{-50}. We can try to make some sense of how small this number is with the observation that 2^{-50} is about equal to the probability that you will be struck by lightning while reading this paragraph. Ain't gonna happen.

The negligible probability that the first fifty flips of a fair coin will all be Heads, let alone that the first one hundred are Heads, simply undermines the credibility of the assumption that the coin is fair. Despite being told the coin is fair, we can't help but consider at least some remote possibility of a mistake: somehow the coin being flipped was not fair but rather was one that had a reasonable chance of flipping one hundred Heads. For example, if a biased coin had probability 0.99 of flipping a Head, then one hundred independent tosses will all come up Heads with probability a little more than one third.

So let's suppose that there are two coins, a fair one and a 0.99 biased one. One of these coins is randomly chosen, with the fair coin hugely favored: the biased coin will be chosen only with the extremely small "struck-by-lightning" probability 2^{-50}. The chosen coin is then flipped one hundred times.

Let E be the event of flipping one hundred heads and H be the event that the chosen coin is the biased one. Now

$$\begin{aligned}
\text{Odds}(H) &= \frac{2^{-50}}{1 - 2^{-50}} \approx 2^{-50}, \\
\text{Bayes-factor}(E, H) &= \frac{\Pr\left[E \mid H\right]}{\Pr\left[E \mid \overline{H}\right]} = \frac{(99/100)^{100}}{2^{-100}} > 0.36 \cdot 2^{100}, \\
\text{Odds}(H \mid E) &= \text{Bayes-factor}(E, H) \cdot \text{Odds}(H) \\
&> 0.36 \cdot 2^{100} \cdot 2^{-50} = 0.36 \cdot 2^{50}.
\end{aligned}$$

So after flipping one hundred Heads, the odds that the biased coin was chosen are overwhelming, namely more than $0.36 \cdot 2^{50}$, in which case the probability that the next flip will be a Head is 0.99. In other words, by assuming some tiny probability that the tossed coin was indeed mistakenly biased toward Heads, we can justify our intuition that after one hundred consecutive Heads, the next flip is very likely to be a Head.

Making an assumption about the probability that some unverified fact is true is known as the *Bayesian* approach to a hypothesis testing problem. By granting a tiny probability that the coin being flipped is biased, the Bayesian approach provides a reasonable justification for estimating that the odds of a Head on the next flip are ninety-nine to one in favor.

18.9.5 Confidence in the Next Flip

If we stick to confidence rather than probability, we don't need to make any Bayesian assumptions about the probability of a fair coin being chosen. We know that if one hundred Heads are flipped, then either the coin is biased, or else the fair coin produced one hundred Heads—something that virtually never happens. This means that when all one hundred flips come up Heads, we can be virtually 100% confident that the coin is biased and therefore 99% confident that the next flip will be a Head.

Problems for Section 18.4

Homework Problems

Problem 18.1.
The Conditional Probability Product Rule for n Events is

Rule.

$$\Pr[E_1 \cap E_2 \cap \ldots \cap E_n] = \Pr[E_1] \cdot \Pr\left[E_2 \mid E_1\right] \cdot \Pr\left[E_3 \mid E_1 \cap E_2\right] \cdots \\ \cdot \Pr\left[E_n \mid E_1 \cap E_2 \cap \ldots \cap E_{n-1}\right] .$$

(a) Restate the Rule without elipses—that is, without “. . . ”—by using “$\bigcap_{i=a}^{b}$” for suitable a, b.

(b) Prove it by induction.

Problems for Section 18.5

Practice Problems

Problem 18.2.
Dirty Harry places two bullets in random chambers of the six-bullet cylinder of his revolver. He gives the cylinder a random spin and says “Feeling lucky?” as he holds the gun against your heart.

(a) What is the probability that you will get shot if he pulls the trigger?

(b) Suppose he pulls the trigger and you don’t get shot. What is the probability that you will get shot if he pulls the trigger a second time?

(c) Suppose you noticed that he placed the two shells next to each other in the cylinder. How does this change the answers to the previous two questions?

Problem 18.3.
State and prove a version of the Law of Total Probability that applies to disjoint events $E_1, \ldots, E_n$ whose union is the whole sample space.

Problem 18.4.
State and prove a version of Bayes Rule that applies to disjoint events $E_1, \ldots, E_n$ whose union is the whole sample space. You may assume the n-event Law of Total Probability, Problem 18.3.

Class Problems

Problem 18.5.
There are two decks of cards. One is complete, but the other is missing the ace of spades. Suppose you pick one of the two decks with equal probability and then select a card from that deck uniformly at random. What is the probability that you picked the complete deck, given that you selected the eight of hearts? Use the four-step method and a tree diagram.

Problem 18.6.
Suppose you have three cards: A♡, A♠ and a jack. From these, you choose a random hand (that is, each card is equally likely to be chosen) of two cards, and let n be the number of aces in your hand. You then randomly pick one of the cards in the hand and reveal it.

(a) Describe a simple probability space (that is, outcomes and their probabilities) for this scenario, and list the outcomes in each of the following events:

1. $[n \geq 1]$, (that is, your hand has an ace in it),
2. A♡ is in your hand,
3. the revealed card is an A♡,
4. the revealed card is an ace.

(b) Then calculate $\Pr\left[n = 2 \mid E\right]$ for E equal to each of the four events in part (a). Notice that most, but *not all*, of these probabilities are equal.

Now suppose you have a deck with d distinct cards, a different kinds of aces (including an A♡), you draw a random hand with h cards, and then reveal a random card from your hand.

(c) Prove that Pr[A♡ is in your hand] $= h/d$.

(d) Prove that

$$\Pr\left[n = 2 \mid \text{A}\heartsuit \text{ is in your hand}\right] = \Pr[n = 2] \cdot \frac{2d}{ah}. \tag{18.8}$$

(e) Conclude that

$$\Pr\left[n = 2 \mid \text{the revealed card is an ace}\right] = \Pr\left[n = 2 \mid \text{A}\heartsuit \text{ is in your hand}\right].$$

Problem 18.7.

A fair six-sided die is repeatedly thrown until a six appears.

A natural sample space modelling this situation is the set of finite strings of integers from one to six that end at the first occurrence of a six. That is, $\mathcal{S} ::= [1..5]^*6$.

For example, 256 is the outcome corresponding to successively throwing a two, a five and a six. The length-one string $6 \in \mathcal{S}$ is the outcome corresponding to six appearing on the first throw.

(a) What should $\Pr[256]$ be defined to be? ...$\Pr[6]$? What about the probability of an arbitrary outcome $s \in \mathcal{S}$?

(b) Verify that $\mathcal{S}$ with the probabilities assigned in part (a) defines a probability space. What does this imply about the possibility of never throwing a six?

For any string $r \in [1..5]^*$, let F_r be the event that values of the initial throws are are the successive elements of r. Let V be the event that all the dice throws are eVen. That is, V is the event $\{2,4\}^*6$ that all throws are twos and fours until the first six.

(c) Suppose t is a string of twos and fours, that is, $t \in \{2,4\}^*$. Explain why

$$\Pr\left[V \mid F_t\right] = \Pr[V]. \tag{18.9}$$

(d) Explain why equation (18.9) implies that for $t \in \{2,4\}^*$,

$$\Pr\left[F_t \mid V\right] = \Pr[F_t].$$

Conclude that

$$\Pr\left[6 \mid V\right] = \frac{2}{3}, \tag{18.10}$$

(e) Given that all throws are even, the only possible first throws are two, four and six. Since the die is fair, these are all equally likely, so the probability $\Pr\left[6 \mid V\right]$ that the first throw is a six must be 1/3, contradicting equation (18.10)! Explain.[6]

(f) Conclude immediately from (18.10) that

$$\Pr[V] = \frac{1}{4}. \tag{18.11}$$

[6] If you're thrown by this, you are not alone. There are several websites devoted to explanations of this seductive problem. In fact, when it came up at the MIT Theory of Computation faculty lunch in April 2018, several attendees confidently defended this mistaken reasoning.

Problem 18.8.
There are three prisoners in a maximum-security prison for fictional villains: the Evil Wizard Voldemort, the Dark Lord Sauron, and Little Bunny Foo-Foo. The parole board has declared that it will release two of the three, chosen uniformly at random, but has not yet released their names. Naturally, Sauron figures that he will be released to his home in Mordor, where the shadows lie, with probability $2/3$.

A guard offers to tell Sauron the name of one of the other prisoners who will be released (either Voldemort or Foo-Foo). If the guard has a choice of naming either Voldemort or Foo-Foo (because both are to be released), he names one of the two with equal probability.

Sauron knows the guard to be a truthful fellow. However, Sauron declines this offer. He reasons that knowing what the guards says will reduce his chances, so he is better off not knowing. For example, if the guard says, “Little Bunny Foo-Foo will be released”, then his own probability of release will drop to $1/2$ because he will then know that either he or Voldemort will also be released, and these two events are equally likely.

Dark Lord Sauron has made a typical mistake when reasoning about conditional probability. Using a tree diagram and the four-step method, **explain his mistake**. What is the probability that Sauron is released given that the guard says Foo-Foo is released?

Hint: Define the events S, F and “F” as follows:

$$\begin{aligned}\text{“}F\text{”} &= \text{Guard says Foo-Foo is released}\\ F &= \text{Foo-Foo is released}\\ S &= \text{Sauron is released}\end{aligned}$$

Problem 18.9.
Every Skywalker serves either the *light side* or the *dark side*.

- The first Skywalker serves the dark side.
- For $n \geq 2$, the n-th Skywalker serves the same side as the $(n-1)$-st Skywalker with probability $1/4$, and the opposite side with probability $3/4$.

Let d_n be the probability that the n-th Skywalker serves the dark side.

(a) Express d_n with a recurrence equation and sufficient base cases.

(b) Derive a simple expression for the generating function $D(x) ::= \sum_1^\infty d_n x^n$.

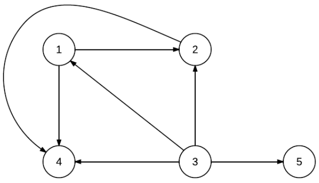


Figure 18.3 The DAG G_0

(c) Give a simple closed formula for d_n.

Problem 18.10. (a) For the directed acyclic graph (DAG) G_0 in Figure 18.3, a minimum-edge DAG with the same walk relation can be obtained by removing some edges. List these edges.

(b) List the vertices in a maximal chain in G_0.

Let G be the simple graph shown in Figure 18.4.

A directed graph $\overrightarrow{G}$ can be randomly constructed from G by assigning a direction to each edge independently with equal likelihood.

(c) What is the probability that $\overrightarrow{G} = G_0$?

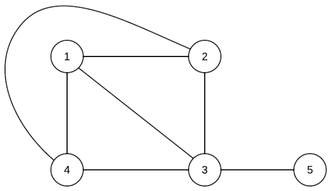


Figure 18.4 Simple graph G

Define the following events with respect to the random graph $\overrightarrow{G}$:

$$
\begin{aligned}
T_1 &::= \text{vertices } 2, 3, 4 \text{ are on a length three directed cycle},\\
T_2 &::= \text{vertices } 1, 3, 4 \text{ are on a length three directed cycle},\\
T_3 &::= \text{vertices } 1, 2, 4 \text{ are on a length three directed cycle},\\
T_4 &::= \text{vertices } 1, 2, 3 \text{ are on a length three directed cycle}.
\end{aligned}
$$

(d) What are $\Pr[T_1], \Pr[T_1 \cap T_2], \Pr[T_1 \cap T_2 \cap T_3]$?

(e) $\overrightarrow{G}$ has the property that if it has a directed cycle, then it has a length three directed cycle. Use this fact to find the probability that $\overrightarrow{G}$ is a DAG.

Homework Problems

Problem 18.11.

There is a subject—naturally not *Math for Computer Science*—in which 10% of the assigned problems contain errors. If you ask a Teaching Assistant (TA) whether a problem has an error, then they will answer correctly 80% of the time, regardless of whether or not a problem has an error. If you ask a lecturer, he will identify whether or not there is an error with only 75% accuracy.

We formulate this as an experiment of choosing one problem randomly and asking a particular TA and Lecturer about it. Define the following events:

$$\begin{aligned}
E &::= [\text{the problem has an error}],\\
T &::= [\text{the TA says the problem has an error}],\\
L &::= [\text{the lecturer says the problem has an error}].
\end{aligned}$$

(a) Translate the description above into a precise set of equations involving conditional probabilities among the events E, T and L.

(b) Suppose you have doubts about a problem and ask a TA about it, and they tell you that the problem is correct. To double-check, you ask a lecturer, who says that the problem has an error. Assuming that the correctness of the lecturer's answer and the TA's answer are independent of each other, regardless of whether there is an error, what is the probability that there is an error in the problem?

(c) Is event T independent of event L (that is, $\Pr\left[T \mid L\right] = \Pr[T]$)? First, give an argument based on intuition, and then calculate both probabilities to verify your intuition.

Problem 18.12.
Suppose you repeatedly flip a fair coin until you see the sequence HTT or HHT. What is the probability you see the sequence HTT first?

Hint: Try to find the probability that HHT comes before HTT conditioning on whether you first toss an H or a T. The answer is not $1/2$.

Problem 18.13.
A 52-card deck is thoroughly shuffled and you are dealt a hand of 13 cards.

(a) If you have one ace, what is the probability that you have a second ace?

(b) If you have the ace of spades, what is the probability that you have a second ace? Remarkably, the answer is different from part (a).

Problem 18.14.
Suppose $\Pr[\cdot] : \mathcal{S} \to [0, 1]$ is a probability function on a sample space $\mathcal{S}$ and let B be an event such that $\Pr[B] > 0$. Define a function $\Pr_B[\cdot]$ on outcomes $\omega \in \mathcal{S}$ by the rule:

$$\Pr_B[\omega] ::= \begin{cases} \Pr[\omega] / \Pr[B] & \text{if } \omega \in B, \\ 0 & \text{if } \omega \notin B. \end{cases} \tag{18.12}$$

(a) Prove that $\Pr_B[\cdot]$ is also a probability function on $\mathcal{S}$ according to Definition 17.5.2.

(b) Prove that

$$\Pr_B[A] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

for all $A \subseteq \mathcal{S}$.

(c) Explain why the Disjoint Sum Rule carries over for conditional probabilities, namely,

$$\Pr\big[C \cup D \mid B\big] = \Pr\big[C \mid B\big] + \Pr\big[D \mid B\big] \qquad (C, D \text{ disjoint}).$$

Give examples of several further such rules.

Problem 18.15.

Professor Meyer has a deck of 52 randomly shuffled playing cards, 26 red, 26 black. He proposes the following game: he will repeatedly draw a card off the top of the deck and turn it face up so that you can see it. At any point while there are still cards left in the deck, you may choose to stop, and he will turn over the next card. If the turned up card is black you win, and otherwise you lose. Either way, the game ends.

Suppose that after drawing off some top cards without stopping, the deck is left with r red cards and b black cards.

(a) Show that if you choose to stop at this point, the probability of winning is $b/(r+b)$.

(b) Prove if you choose *not* to stop at this point, the probability of winning is still $b/(r+b)$, regardless of your stopping strategy for the rest of the game.

Hint: Induction on $r + b$.

Exam Problems

Problem 18.16.

Sally Smart just graduated from high school. She was accepted to three reputable colleges.

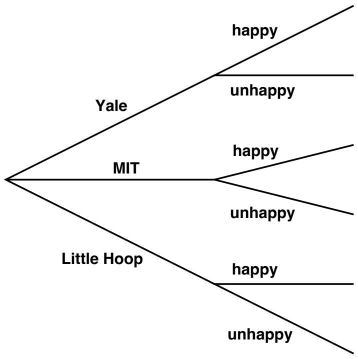
- With probability $4/12$, she attends Yale.
- With probability $5/12$, she attends MIT.

- With probability $3/12$, she attends Little Hoop Community College.

Sally is either happy or unhappy in college.

- If she attends Yale, she is happy with probability $4/12$.
- If she attends MIT, she is happy with probability $7/12$.
- If she attends Little Hoop, she is happy with probability $11/12$.

(a) A tree diagram to help Sally project her chance at happiness is shown below. On the diagram, fill in the edge probabilities, and at each leaf write the probability of the corresponding outcome.



(b) What is the probability that Sally is happy in college?

(c) What is the probability that Sally attends Yale, given that she is happy in college?

(d) Show that the event that Sally attends Yale **is not** independent of the event that she is happy.

(e) Show that the event that Sally attends MIT **is** independent of the event that she is happy.

Problem 18.17.

Here's a variation of Monty Hall's game: the contestant still picks one of three doors, with a prize randomly placed behind one door and goats behind the other two. But now, instead of always opening a door to reveal a goat, Monty instructs

Carol to *randomly* open one of the two doors that the contestant hasn't picked. This means she may reveal a goat, or she may reveal the prize. If she reveals the prize, then the entire game is *restarted*, that is, the prize is again randomly placed behind some door, the contestant again picks a door, and so on until Carol finally picks a door with a goat behind it. Then the contestant can choose to *stick* with his original choice of door or *switch* to the other unopened door. He wins if the prize is behind the door he finally chooses.

To analyze this setup, we define two events:

GP**:** The event that the contestant **g**uesses the door with the **p**rize behind it on his first guess.

OP**:** The event that the game is restarted at least once. Another way to describe this is as the event that the door Carol first **o**pens has a **p**rize behind it.

Give the values of the following probabilities:

(a) $\Pr[GP]$

(b) $\Pr\left[OP \mid \overline{GP}\right]$

(c) $\Pr[OP]$

(d) the probability that the game will continue forever

(e) When Carol finally picks the goat, the contestant has the choice of sticking or switching. Let's say that the contestant adopts the strategy of sticking. Let W be the event that the contestant wins with this strategy, and let $w ::= \Pr[W]$. Express the following conditional probabilities as simple closed forms in terms of w.

i) $\Pr\left[W \mid GP\right]$

ii) $\Pr\left[W \mid \overline{GP} \cap OP\right]$

iii) $\Pr\left[W \mid \overline{GP} \cap \overline{OP}\right]$

(f) What is the value of $\Pr[W]$?

(g) For any final outcome where the contestant wins with a "stick" strategy, he would lose if he had used a "switch" strategy, and vice versa. In the original Monty Hall game, we concluded immediately that the probability that he would win with a "switch" strategy was $1 - \Pr[W]$. Why isn't this conclusion quite as obvious for this new, restartable game? Is this conclusion still sound? Briefly explain.

Problem 18.18.
There are two decks of cards, the red deck and the blue deck. They differ slightly in a way that makes drawing the eight of hearts slightly more likely from the red deck than from the blue deck.

One of the decks is randomly chosen and hidden in a box. You reach in the box and randomly pick a card that turns out to be the eight of hearts. You believe intuitively that this makes the red deck more likely to be in the box than the blue deck.

Your intuitive judgment about the red deck can be formalized and verified using some inequalities between probabilities and conditional probabilities involving the events

R ::= **R**ed deck is in the box,
B ::= **B**lue deck is in the box,
E ::= **E**ight of hearts is picked from the deck in the box.

(a) State an inequality between probabilities and/or conditional probabilities that formalizes the assertion, "picking the eight of hearts from the red deck is more likely than from the blue deck."

(b) State a similar inequality that formalizes the assertion "picking the eight of hearts from the deck in the box makes the red deck more likely to be in the box than the blue deck."

(c) Assuming that initially each deck is equally likely to be the one in the box, prove that the inequality of part (a) implies the inequality of part (b).

(d) Suppose you couldn't be sure that the red deck and blue deck initially were equally likely to be in the box. Could you still conclude that picking the eight of hearts from the deck in the box makes the red deck more likely to be in the box than the blue deck? Briefly explain.

Problem 18.19.
A flip of Coin 1 is x times as likely to come up Heads as a flip of Coin 2. A biased random choice of one of these coins will be made, where the probability of choosing Coin 1 is w times that of Coin 2.

(a) Restate the information above as equations between conditional probabilities involving the events

$$\begin{aligned} C1 &::= \text{Coin 1 was chosen,}\\ C2 &::= \text{Coin 2 was chosen,}\\ H &::= \text{the chosen coin came up Heads.} \end{aligned}$$

(b) State an inequality involving conditional probabilities of the above events that formalizes the assertion “Given that the chosen coin came up Heads, the chosen coin is more likely to have been Coin 1 than Coin 2.”

(c) Prove that, given that the chosen coin came up Heads, the chosen coin is more likely to have been Coin 1 than Coin 2 iff

$$wx > 1.$$

Problem 18.20.
There is an unpleasant, degenerative disease called Beaver Fever which causes people to tell math jokes unrelentingly in social settings, believing other people will think they’re funny. Fortunately, Beaver Fever is rare, afflicting only about 1 in 1000 people. Doctor Meyer has a fairly reliable diagnostic test to determine who is going to suffer from this disease:

- If a person will suffer from Beaver Fever, the probability that Dr. Meyer diagnoses this is 0.99.
- If a person will not suffer from Beaver Fever, the probability that Dr. Meyer diagnoses this is 0.97.

Let B be the event that a randomly chosen person will suffer **B**eaver Fever, and Y be the event that Dr. Meyer’s diagnosis is “**Y**es, this person will suffer from Beaver Fever,” with $\overline{B}$ and $\overline{Y}$ being the complements of these events.

(a) The description above explicitly gives the values of the following quantities. What are their values?

$$\Pr[B] \qquad \Pr\left[Y \mid B\right] \qquad \Pr\left[\overline{Y} \mid \overline{B}\right]$$

(b) Write formulas for $\Pr[\overline{B}]$ and $\Pr\left[Y \mid \overline{B}\right]$ solely in terms of the explicitly given quantities in part (a)—literally use their expressions, not their numeric values.

(c) Write a formula for the probability that Dr. Meyer says a person will suffer from Beaver Fever solely in terms of $\Pr[B]$, $\Pr[\overline{B}]$, $\Pr\left[Y \mid B\right]$ and $\Pr\left[Y \mid \overline{B}\right]$.

(d) Write a formula solely in terms of the expressions given in part (a) for the probability that a person will suffer Beaver Fever given that Doctor Meyer says they will. Then calculate the numerical value of the formula.

Suppose there was a vaccine to prevent Beaver Fever, but the vaccine was expensive or slightly risky itself. If you were sure you were going to suffer from Beaver Fever, getting vaccinated would be worthwhile, but even if Dr. Meyer diagnosed you as a future sufferer of Beaver Fever, the probability you actually will suffer Beaver Fever remains low (about 1/32 by part (d)).

In this case, you might sensibly decide not to be vaccinated—after all, Beaver Fever is not *that* bad an affliction. So the diagnostic test serves no purpose in your case. You may as well not have bothered to get diagnosed. Even so, the test may be useful:

(e) Suppose Dr. Meyer had enough vaccine to treat 2% of the population. If he randomly chose people to vaccinate, he could expect to vaccinate only 2% of the people who needed it. But by testing everyone and only vaccinating those diagnosed as future sufferers, he can expect to vaccinate a much larger fraction people who were going to suffer from Beaver Fever. Estimate this fraction.

Problem 18.21.

Suppose that *Let's Make a Deal* is played according to slightly different rules and with a red goat and a blue goat. There are three doors, with a prize hidden behind one of them and one goat behind each of the others. No doors are opened until the contestant makes a final choice to stick or switch. The contestant is allowed to pick a door and ask a certain question that the host then answers honestly. The contestant may then stick with their chosen door, or switch to either of the other doors.

(a) If the contestant asks "is there is a goat behind one of the unchosen doors?" and the host answers "yes," is the contestant more likely to win the prize if they stick, switch, or does it not matter? Clearly identify the probability space of outcomes and their probabilities you use to model this situation. What is the contestant's probability of winning if he uses the best strategy?

(b) If the contestant asks "is the *red* goat behind one of the unchosen doors?" and the host answers "yes," is the contestant more likely to win the prize if they stick, switch, or does it not matter? Clearly identify the probability space of outcomes and their probabilities you use to model this situation. What is the contestant's probability of winning if he uses the best strategy?

Problem 18.22.
You are organizing a neighborhood census and instruct your census takers to knock on doors and note the gender of any child that answers the knock. Assume that there are two children in every household, that a random child is equally likely to be a girl or a boy, and that the two children in a household are equally likely to be the one that opens the door.

A sample space for this experiment has outcomes that are triples whose first element is either B or G for the gender of the elder child, whose second element is either B or G for the gender of the younger child, and whose third element is E or Y indicating whether the *e*lder child or *y*ounger child opened the door. For example, $(\texttt{B}, \texttt{G}, \texttt{Y})$ is the outcome that the elder child is a boy, the younger child is a girl, and the girl opened the door.

(a) Let O be the event that a girl opened the door, and let T be the event that the household has two girls. List the outcomes in O and the outcomes in T.

(b) What is the probability $\Pr[T \mid O]$, that both children are girls, given that a girl opened the door?

(c) What mistake is made in the following argument? (For example, perhaps there's an arithmetic mistake [where?], or an unjustified assumption [what?], or some other error entirely.) Please identify and explain the error in detail.

> If a girl opens the door, then we know that there is at least one girl in the household. The probability that there is at least one girl is
>
> $$1 - \Pr[\text{both children are boys}] = 1 - (1/2 \times 1/2) = 3/4.$$
>
> So,
>
> $$\begin{aligned}\Pr\left[T \mid \text{there is at least one girl in the household}\right] \\ = \frac{\Pr[T \cap \text{there is at least one girl in the household}]}{\Pr[\text{there is at least one girl in the household}]} \\ = \frac{\Pr[T]}{\Pr[\text{there is at least one girl in the household}]} \\ = (1/4)/(3/4) = 1/3.\end{aligned}$$
>
> Therefore, given that a girl opened the door, the probability that there are two girls in the household is 1/3.

Problem 18.23.
A guard is going to release exactly two of the three prisoners, Sauron, Voldemort, and Bunny Foo Foo, and he’s equally likely to release any set of two prisoners.

(a) What is the probability that Voldemort will be released?

The guard will truthfully tell Voldemort the name of one of the prisoners to be released. We’re interested in the following events:

V**:** *V*oldemort is released.

“F”: The guard tells Voldemort that *F*oo Foo will be released.

“S”: The guard tells Voldemort that *S*auron will be released.

The guard has two rules for choosing whom he names:

- never say that Voldemort will be released,
- if both Foo Foo and Sauron are getting released, say “Foo Foo.”

(b) What is $\Pr[V \mid \text{“}F\text{”}]$?

(c) What is $\Pr[V \mid \text{“}S\text{”}]$?

(d) Show how to use the Law of Total Probability to combine your answers to parts (b) and (c) to verify that the result matches the answer to part (a).

Problem 18.24.
We are interested in paths in the plane starting at (0, 0) that go one unit right or one unit up at each step. To model this, we use a state machine whose states are $\mathbb{N} \times \mathbb{N}$, whose start state is (0, 0), and whose transitions are

$$(x, y) \to (x + 1, y),$$
$$(x, y) \to (x, y + 1).$$

(a) How many length n paths are there starting from the origin?

(b) How many states are reachable in exactly n steps?

(c) How many states are reachable in at most n steps?

(d) If transitions occur independently at random, going right with probability p and up with probability $q ::= 1 - p$ at each step, what is the probability of reaching position (x, y)?

(e) What is the probability of reaching state (x, y) *given* that the path to (x, y) reached (m, n) before getting to (x, y)?

(f) Show that the probability that a path ending at (x, y) went through (m, n) is the same for all p.

Problems for Section 18.6

Practice Problems

Problem 18.25.
Define the events A, F_{EE}, F_{CS}, M_{EE}, and M_{CS} as in Section 18.6.

In these terms, the plaintiff in a discrimination suit against a university makes the argument that in both departments, the probability that a female is admitted is less than the probability for a male. That is,

$$\Pr[A \mid F_{EE}] < \Pr[A \mid M_{EE}] \quad \text{and} \tag{18.13}$$

$$\Pr[A \mid F_{CS}] < \Pr[A \mid M_{CS}]. \tag{18.14}$$

The university's defence attorneys retort that *overall*, a female applicant is *more* likely to be admitted than a male, namely, that

$$\Pr[A \mid F_{EE} \cup F_{CS}] > \Pr[A \mid M_{EE} \cup M_{CS}]. \tag{18.15}$$

The judge then interrupts the trial and calls the plaintiff and defence attorneys to a conference in his office to resolve what he thinks are contradictory statements of facts about the admission data. The judge points out that:

$$\begin{aligned}
&\Pr[A \mid F_{EE} \cup F_{CS}] \\
&= \Pr[A \mid F_{EE}] + \Pr[A \mid F_{CS}] && \text{(because } F_{EE} \text{ and } F_{CS} \text{ are disjoint)} \\
&< \Pr[A \mid M_{EE}] + \Pr[A \mid M_{CS}] && \text{(by (18.13) and (18.14))} \\
&= \Pr[A \mid M_{EE} \cup M_{CS}] && \text{(because } M_{EE} \text{ and } M_{CS} \text{ are disjoint)}
\end{aligned}$$

so

$$\Pr[A \mid F_{EE} \cup F_{CS}] < \Pr[A \mid M_{EE} \cup M_{CS}],$$

which directly contradicts the university's position (18.15)!

Of course the judge is mistaken; an example where the plaintiff and defence assertions are all true appears in Section 18.6. What is the mistake in the judge's proof?

Problems for Section 18.7

Practice Problems

Problem 18.26.
Outside of their hum-drum duties as Math for Computer Science Teaching Assistants, Oscar is trying to learn to levitate using only intense concentration and Liz is trying to become the world champion flaming torch juggler. Suppose that Oscar's probability of success is $1/6$, Liz's chance of success is $1/4$, and these two events are independent.

(a) If at least one of them succeeds, what is the probability that Oscar learns to levitate?

(b) If at most one of them succeeds, what is the probability that Liz becomes the world flaming torch juggler champion?

(c) If exactly one of them succeeds, what is the probability that it is Oscar?

Problem 18.27.
What is the smallest size sample space in which there are two independent events, neither of which has probability zero or probability one? Explain.

Problem 18.28.
Give examples of event A, B, E such that

(a) A and B are independent, and are also conditionally independent given E, but are not conditionally independent given $\overline{E}$. That is,

$$\begin{aligned}
\Pr[A \cap B] &= \Pr[A] \Pr[B], \\
\Pr\left[A \cap B \mid E\right] &= \Pr\left[A \mid E\right] \Pr\left[B \mid E\right], \\
\Pr\left[A \cap B \mid \overline{E}\right] &\neq \Pr\left[A \mid \overline{E}\right] \Pr\left[B \mid \overline{E}\right].
\end{aligned}$$

Hint: Let $\mathcal{S} = \{1, 2, 3, 4\}$.

(b) A and B are conditionally independent given E, or given $\overline{E}$, but are not independent. That is,

$$\begin{aligned}
\Pr\left[A \cap B \mid E\right] &= \Pr\left[A \mid E\right] \Pr\left[B \mid E\right], \\
\Pr\left[A \cap B \mid \overline{E}\right] &= \Pr\left[A \mid \overline{E}\right] \Pr\left[B \mid \overline{E}\right], \\
\Pr[A \cap B] &\neq \Pr[A] \Pr[B].
\end{aligned}$$

Hint: Let $\mathcal{S} = \{1, 2, 3, 4, 5\}$.

Class Problems

Problem 18.29.

Event E is *evidence in favor* of event H when $\Pr\left[H \mid E\right] > \Pr[H]$, and it is *evidence against* H when $\Pr\left[H \mid E\right] < \Pr[H]$.

(a) Give an example of events A, B, H such that A and B are independent, both are evidence for H, but $A \cup B$ is evidence against H.

Hint: Let $\mathcal{S} = [1..8]$

(b) Prove E is evidence in favor of H iff $\overline{E}$ is evidence against H.

Problem 18.30.

Let G be a simple graph with n vertices. Let "$A(u, v)$" mean that vertices u and v are adjacent, and let "$W(u, v)$" mean that there is a length-two walk between u and v.

(a) Explain why $W(u, u)$ holds iff $\exists v.\ A(u, v)$.

(b) Write a predicate-logic formula defining $W(u, v)$ in terms of the predicate $A(.,.)$ when $u \neq v$.

There are $e ::= \binom{n}{2}$ possible edges between the n vertices of G. Suppose the actual edges of $E(G)$ are chosen with randomly from this set of e possible edges. Each edge is chosen with probability p, and the choices are mutually independent.

(c) Write a simple formula in terms of p, e and k for $\Pr[|E(G)| = k]$.

(d) Write a simple formula in terms of p and n for $\Pr[W(u, u)]$.

Let w, x, y and z be four distinct vertices.

Because edges are chosen mutually independently, if the edges that one event depends on are disjoint from the edges that another event depends on, then the two events will be mutually independent. For example, the events

$$A(w, y) \text{ AND } A(y, x)$$

and

$$A(w, z) \text{ AND } A(z, x)$$

are independent since the first event dependes on $\{\langle w\text{—}y\rangle, \langle y\text{—}x\rangle\}$, while the second event depends on $\{\langle w\text{—}z\rangle, \langle z\text{—}x\rangle\}$.

(e) Let

$$r ::= \Pr[\text{NOT}(W(w, x))], \tag{18.16}$$

where w and x are distinct vertices. Write a simple formula for r in terms of n and p.

Hint: Different length-two paths between x and y don't share any edges.

(f) Vertices x and y being on a three-cycle can be expressed simply as

$$A(x, y) \text{ AND } W(x, y).$$

Write a simple expression in terms of p and r for the probability that x and y lie on a three-cycle in G.

(g) Show that $W(w, x)$ and $W(y, z)$ may not be independent events. *Hint:* Just consider the case that $V(G) = \{w, x, y, z\}$ and $p = 1/2$.

Exam Problems

Problem 18.31. **(a)** Show that any total, symmetric, transitive relation is reflexive.

(b) Conclude that there are events A, B, C such that A is independent of B, B is independent of C, but C is not independent of A.

Problems for Section 18.8

Practice Problems

Problem 18.32.
Suppose A, B and C are mutually independent events, what about $A \cap B$ and $B \cup C$?

Class Problems

Problem 18.33.
Suppose you flip three fair, mutually independent coins. Define the following events:

- Let A be the event that *the first* coin is heads.
- Let B be the event that *the second* coin is heads.
- Let C be the event that *the third* coin is heads.
- Let D be the event that *an even number of* coins are heads.

(a) Use the four step method to determine the probability space for this experiment and the probability of each of A, B, C, D.

(b) Show that these events are not mutually independent.

(c) Show that they are 3-way independent.

Problem 18.34.
Let A, B, C be events. For each of the following statements, prove it or give a counterexample.

(a) If A is independent of B, then A is also independent of $\overline{B}$.

(b) If A is independent of B, and A is independent of C, then A is independent of $B \cap C$.

Hint: Choose A, B, C pairwise but not 3-way independent.

(c) If A is independent of B, and A is independent of C, then A is independent of $B \cup C$.

Hint: Part (b).

(d) If A is independent of B, and A is independent of C, and A is independent of $B \cap C$, then A is independent of $B \cup C$.

Problem 18.35.
Let A, B, C, D be events. Describe counterexamples showing that the following claims are false.

(a)

False Claim. *If A and B are independent given C, and are also independent given D, then A and B are independent given $C \cup D$.*

(b)

False Claim. *If A and B are independent given C, and are also independent given D, then A and B are independent given $C \cap D$.*

Hint: Choose A, B, C, D 3-way but not 4-way independent.

so A and B are not independent given $C \cap D$.

Homework Problems

Problem 18.36.
Describe events A, B and C that:

- satisfy the "product rule," namely,

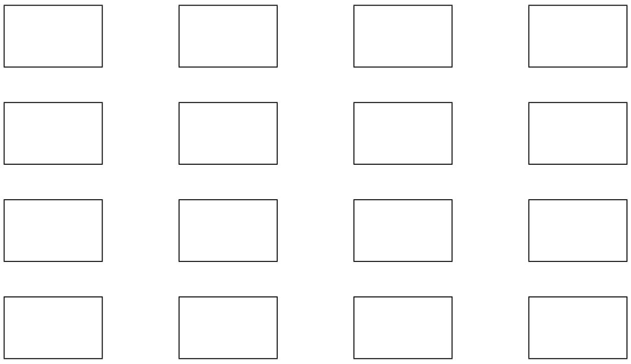
$$\Pr[A \cap B \cap C] = \Pr[A] \cdot \Pr[B] \cdot \Pr[C],$$

- and additionally, no two out of the three events are independent.

Hint: Choose A, B, C events over the uniform probability space on [1..6].

Exam Problems

Problem 18.37.
A classroom has sixteen desks in a 4×4 arrangement as shown below.



If two desks are next to each other, vertically or horizontally, they are called an *adjacent pair*. So there are three horizontally adjacent pairs in each row, for a total of twelve horizontally adjacent pairs. Likewise, there are twelve vertically adjacent pairs.

Boys and girls are assigned to desks mutually independently, with probability $p > 0$ of a desk being occupied by a boy and probability $q ::= 1 - p > 0$ of being occupied by a girl. An adjacent pair D of desks is said to have a *flirtation* when there is a boy at one desk and a girl at the other desk. Let F_D be the event that D has a flirtation.

(a) Different pairs D and E of adjacent desks are said to *overlap* when they share a desk. For example, the first and second pairs in each row overlap, and so do the second and third pairs, but the first and third pairs do not overlap. Prove that if D and E overlap, then F_D and F_E are independent events iff $p = q$.

(b) Find four pairs of desks D_1, D_2, D_3, D_4 and explain why $F_{D_1}, F_{D_2}, F_{D_3}, F_{D_4}$ are *not* mutually independent (even if $p = q = 1/2$).

Problems for Section 18.9

Problem 18.38.
An *International Journal of Pharmacological Testing* has a policy of publishing drug trial results only if the conclusion holds at the 95% confidence level. The editors and reviewers always carefully check that any results they publish came from a drug trial that genuinely deserved this level of confidence. They are also careful to check that trials whose results they publish have been conducted independently of each other.

The editors of the Journal reason that under this policy, their readership can be confident that at most 5% of the published studies will be mistaken. Later, the editors are embarrassed—and astonished—to learn that *every one* of the 20 drug trial results they published during the year was wrong. The editors thought that because the trials were conducted independently, the probability of publishing 20 wrong results was negligible, namely, $(1/20)^{20} < 10^{-25}$.

Write a brief explanation to these befuddled editors explaining what's wrong with their reasoning and how it could be that all 20 published studies were wrong.

Hint: xkcd comic: "significant" xkcd.com/882/

Practice Problems

Problem 18.39.
A somewhat reliable allergy test has the following properties:

- If you are allergic, there is a 10% chance that the test will say you are not.
- If you are not allergic, there is a 5% chance that the test will say you are.

(a) The test results are correct at what confidence level?

(b) What is the Bayes factor for being allergic when the test diagnoses a person as allergic?

(c) What can you conclude about the odds of a random person being allergic given that the test diagnoses them as allergic? Can you determine if the odds are better than even?

Suppose that your doctor tells you that because the test diagnosed you as allergic, and about 25% of people are allergic, the odds are six to one that you are allergic.

(d) How would your doctor calculate these odds of being allergic based on what's known about the allergy test?

(e) Another doctor reviews your test results and medical record and says your odds of being allergic are really much higher, namely thirty-six to one. Briefly explain how two conscientious doctors could disagree so much. Is there a way you could determine your actual odds of being allergic?

19 Random Variables

Thus far, we have focused on probabilities of events. For example, we computed the probability that you win the Monty Hall game or that you have a rare medical condition given that you tested positive. But, in many cases we would like to know more. For example, *how many* contestants must play the Monty Hall game until one of them finally wins? *How long* will this condition last? *How much* will I lose gambling with strange dice all night? To answer such questions, we need to work with random variables.

19.1 Random Variable Examples

Definition 19.1.1. A *random variable* R on a probability space is a total function whose domain is the sample space.

The codomain of R can be anything, but will usually be a subset of the real numbers. Notice that the name "random variable" is a misnomer; random variables are actually functions.

For example, suppose we toss three independent, unbiased coins. Let C be the number of heads that appear. Let $M = 1$ if the three coins come up all heads or all tails, and let $M = 0$ otherwise. Now every outcome of the three coin flips uniquely determines the values of C and M. For example, if we flip heads, tails, heads, then $C = 2$ and $M = 0$. If we flip tails, tails, tails, then $C = 0$ and $M = 1$. In effect, C counts the number of heads, and M indicates whether all the coins match.

Since each outcome uniquely determines C and M, we can regard them as functions mapping outcomes to numbers. For this experiment, the sample space is:

$$\mathcal{S} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Now C is a function that maps each outcome in the sample space to a number as follows:

$$\begin{array}{rclcrcl}
C(HHH) & = & 3 & \qquad & C(THH) & = & 2 \\
C(HHT) & = & 2 & & C(THT) & = & 1 \\
C(HTH) & = & 2 & & C(TTH) & = & 1 \\
C(HTT) & = & 1 & & C(TTT) & = & 0.
\end{array}$$

Similarly, M is a function mapping each outcome another way:

$$\begin{array}{rclcrcl} M(HHH) & = & 1 & \qquad & M(THH) & = & 0 \\ M(HHT) & = & 0 & & M(THT) & = & 0 \\ M(HTH) & = & 0 & & M(TTH) & = & 0 \\ M(HTT) & = & 0 & & M(TTT) & = & 1. \end{array}$$

So C and M are random variables.

19.1.1 Indicator Random Variables

An *indicator random variable* is a random variable that maps every outcome to either 0 or 1. Indicator random variables are also called *Bernoulli variables*. The random variable M is an example. If all three coins match, then $M = 1$; otherwise, $M = 0$.

Indicator random variables are closely related to events. In particular, an indicator random variable partitions the sample space into those outcomes mapped to 1 and those outcomes mapped to 0. For example, the indicator M partitions the sample space into two blocks as follows:

$$\underbrace{HHH \quad TTT}_{M=1} \quad \underbrace{HHT \quad HTH \quad HTT \quad THH \quad THT \quad TTH}_{M=0}.$$

In the same way, an event E partitions the sample space into those outcomes in E and those not in E. So E is naturally associated with an indicator random variable, I_E, where $I_E(\omega) = 1$ for outcomes $\omega \in E$ and $I_E(\omega) = 0$ for outcomes $\omega \notin E$. This means that event E is the same as the event $[I_E = 1]$. For example the variable M above is really just the indicator variable I_E, where E is the event that all three coins match.

19.1.2 Random Variables and Events

There is a strong relationship between events and more general random variables as well. A random variable that takes on several values partitions the sample space into several blocks. For example, C partitions the sample space as follows:

$$\underbrace{TTT}_{C=0} \quad \underbrace{TTH \quad THT \quad HTT}_{C=1} \quad \underbrace{THH \quad HTH \quad HHT}_{C=2} \quad \underbrace{HHH}_{C=3}.$$

Each block is a subset of the sample space and is therefore an event. So the assertion that $C = 2$ defines the event

$$[C = 2] = \{THH, HTH, HHT\},$$

and this event has probability

$$\Pr[C = 2] = \Pr[THH] + \Pr[HTH] + \Pr[HHT] = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = 3/8.$$

Likewise $[M = 1]$ is the event $\{TTT, HHH\}$ and has probability $1/4$.

More generally, any assertion about the values of random variables defines an event. For example, the assertion that $C \le 1$ defines

$$[C \le 1] = \{TTT, TTH, THT, HTT\},$$

and so $\Pr[C \le 1] = 1/2$.

Another example is the assertion that $C \cdot M$ is an odd number. If you think about it for a minute, you'll realize that this is an obscure way of saying that all three coins came up heads, namely,

$$[C \cdot M \text{ is odd}] = \{HHH\}.$$

19.2 Independence

The notion of independence carries over from events to random variables as well. Random variables R_1 and R_2 are *independent* iff for all x_1, x_2, the two events

$$[R_1 = x_1] \quad \text{and} \quad [R_2 = x_2]$$

are independent.

For example, are C and M independent? Intuitively, the answer should be "no." The number of heads C completely determines whether all three coins match; that is, whether $M = 1$. But, to verify this intuition, we must find some $x_1, x_2 \in \mathbb{R}$ such that:

$$\Pr[C = x_1 \text{ AND } M = x_2] \neq \Pr[C = x_1] \cdot \Pr[M = x_2].$$

One appropriate choice of values is $x_1 = 2$ and $x_2 = 1$. In this case, we have:

$$\Pr[C = 2 \text{ AND } M = 1] = 0 \neq \frac{1}{4} \cdot \frac{3}{8} = \Pr[M = 1] \cdot \Pr[C = 2].$$

The first probability is zero because we never have exactly two heads ($C = 2$) when all three coins match ($M = 1$). The other two probabilities were computed earlier.

On the other hand, let H_1 be the indicator variable for the event that the first flip is a Head, so

$$[H_1 = 1] = \{HHH, HTH, HHT, HTT\}.$$

Then H_1 is independent of M, since

$$\begin{aligned}\Pr[M = 1] = 1/4 = \Pr\big[M = 1 \mid H_1 = 1\big] = \Pr\big[M = 1 \mid H_1 = 0\big]\\ \Pr[M = 0] = 3/4 = \Pr\big[M = 0 \mid H_1 = 1\big] = \Pr\big[M = 0 \mid H_1 = 0\big]\end{aligned}$$

This example is an instance of:

Lemma 19.2.1. *Two events are independent iff their indicator variables are independent.*

The simple proof is left to Problem 19.1.

Intuitively, the independence of two random variables means that knowing some information about one variable doesn’t provide any information about the other one. We can formalize what “some information” about a variable R is by defining it to be the value of some quantity that depends on R. This intuitive property of independence then simply means that functions of independent variables are also independent:

Lemma 19.2.2. *Let R and S be independent random variables, and f and g be functions such that* $\text{domain}(f) = \text{codomain}(R)$ *and* $\text{domain}(g) = \text{codomain}(S)$. *Then $f(R)$ and $g(S)$ are independent random variables.*

The proof is another simple exercise left to Problem 19.36.

As with events, the notion of independence generalizes to more than two random variables.

Definition 19.2.3. Random variables $R_1, R_2, \dots, R_n$ are *mutually independent* iff for all $x_1, x_2, \dots, x_n$, the n events

$$[R_1 = x_1], [R_2 = x_2], \dots, [R_n = x_n]$$

are mutually independent. They are *k-way independent* iff every subset of k of them are mutually independent.

Lemmas 19.2.1 and 19.2.2 both extend straightforwardly to k-way independent variables.

19.3 Distribution Functions

A random variable maps outcomes to values. The probability density function, $\text{PDF}_R(x)$, of a random variable R measures the probability that R takes the value x, and the closely related cumulative distribution function $\text{CDF}_R(x)$ measures the probability that $R \le x$. Random variables that show up for different spaces of outcomes often wind up behaving in much the same way because they have the same probability of taking different values, that is, because they have the same pdf/cdf.

Definition 19.3.1. Let R be a random variable with codomain V. The *probability density function* of R is a function $\text{PDF}_R : V \to [0, 1]$ defined by:

$$\text{PDF}_R(x) ::= \begin{cases} \Pr[R = x] & \text{if } x \in \text{range}(R), \\ 0 & \text{if } x \notin \text{range}(R). \end{cases}$$

If the codomain is a subset of the real numbers, then the *cumulative distribution function* is the function $\text{CDF}_R : \mathbb{R} \to [0, 1]$ defined by:

$$\text{CDF}_R(x) ::= \Pr[R \le x].$$

A consequence of this definition is that

$$\sum_{x \in \text{range}(R)} \text{PDF}_R(x) = 1.$$

This is because R has a value for each outcome, so summing the probabilities over all outcomes is the same as summing over the probabilities of each value in the range of R.

As an example, suppose that you roll two unbiased, independent, 6-sided dice. Let T be the random variable that equals the sum of the two rolls. This random variable takes on values in the set $V = \{2, 3, \dots, 12\}$. A plot of the probability density function for T is shown in Figure 19.1. The lump in the middle indicates that sums close to seven are the most likely. The total area of all the rectangles is 1 since the dice must take on exactly one of the sums in $V = \{2, 3, \dots, 12\}$.

The cumulative distribution function for T is shown in Figure 19.2: The height of the ith bar in the cumulative distribution function is equal to the *sum* of the heights of the leftmost i bars in the probability density function. This follows from the definitions of pdf and cdf:

$$\text{CDF}_R(x) = \Pr[R \le x] = \sum_{y \le x} \Pr[R = y] = \sum_{y \le x} \text{PDF}_R(y).$$

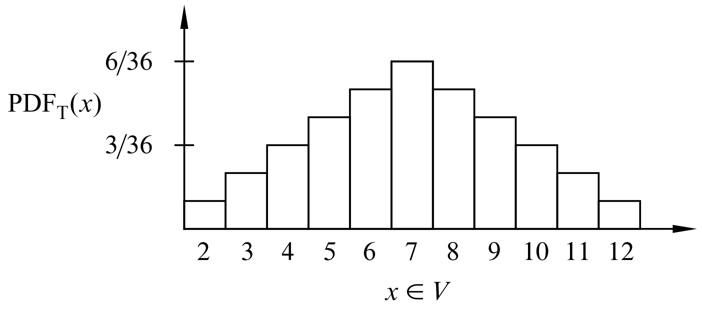


Figure 19.1 The probability density function for the sum of two 6-sided dice.

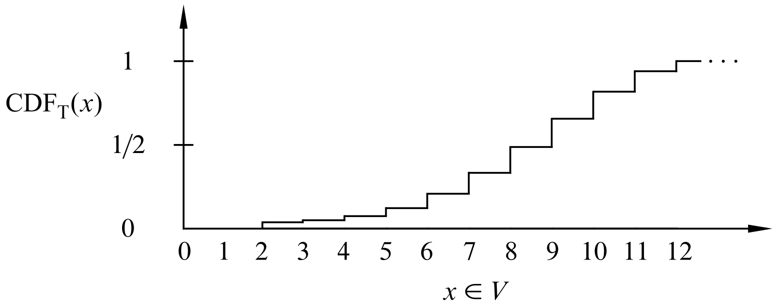


Figure 19.2 The cumulative distribution function for the sum of two 6-sided dice.

It also follows from the definition that

$$\lim_{x \to \infty} \text{CDF}_R(x) = 1 \text{ and } \lim_{x \to -\infty} \text{CDF}_R(x) = 0.$$

Both PDF_R and CDF_R capture the same information about R, so take your choice. The key point here is that neither the probability density function nor the cumulative distribution function involves the sample space of an experiment.

One of the really interesting things about density functions and distribution functions is that many random variables turn out to have the *same* pdf and cdf. In other words, even though R and S are different random variables on different probability spaces, it is often the case that

$$\text{PDF}_R = \text{PDF}_S.$$

In fact, some pdf’s are so common that they are given special names. For example, the most important distributions in computer science arguably are the *Bernoulli distribution*, the *Uniform distribution*, the *Binomial distribution*, and the *Geometric distribution*. We look more closely at these common distributions in the next several sections.

19.3.1 Bernoulli Distributions

A Bernoulli distribution is the distribution function for a Bernoulli variable. Specifically, the *Bernoulli distribution* has a probability density function of the form $f_p : \{0, 1\} \to [0, 1]$ where

$$\begin{aligned} f_p(0) &= p, \quad \text{and} \\ f_p(1) &= q, \end{aligned}$$

for some $p \in [0, 1]$ with $q ::= 1 - p$. The corresponding cumulative distribution function is $F_p : \mathbb{R} \to [0, 1]$ where

$$F_p(x) ::= \begin{cases} 0 & \text{if } x < 0 \\ p & \text{if } 0 \le x < 1 \\ 1 & \text{if } 1 \le x. \end{cases}$$

19.3.2 Uniform Distributions

A random variable that takes on each possible value in its codomain with the same probability is said to be *uniform*. If the codomain V has n elements, then the *uniform distribution* has a pdf of the form

$$f : V \to [0, 1]$$

where

$$f(v) = \frac{1}{n}$$

for all $v \in V$.

If the elements of V in increasing order are $a_1, a_2, \dots, a_n$, then the cumulative distribution function would be $F : \mathbb{R} \to [0, 1]$ where

$$F(x) ::= \begin{cases} 0 & \text{if } x < a_1 \\ k/n & \text{if } a_k \leq x < a_{k+1} \text{ for } 1 \leq k < n \\ 1 & \text{if } a_n \leq x. \end{cases}$$

Uniform distributions come up all the time. For example, the number rolled on a fair die is uniform on the set $\{1, 2, \dots, 6\}$. An indicator variable is uniform when its pdf is $f_{1/2}$.

19.3.3 The Numbers Game

Enough definitions—let's play a game! We have two envelopes. Each contains an integer in the range $0, 1, \dots, 100$, and the numbers are distinct. To win the game, you must determine which envelope contains the larger number. To give you a fighting chance, we'll let you peek at the number in one envelope selected at random. Can you devise a strategy that gives you a better than 50% chance of winning?

For example, you could just pick an envelope at random and guess that it contains the larger number. But this strategy wins only 50% of the time. Your challenge is to do better.

So you might try to be more clever. Suppose you peek in one envelope and see the number 12. Since 12 is a small number, you might guess that the number in the other envelope is larger. But perhaps we've been tricky and put small numbers in *both* envelopes. Then your guess might not be so good!

An important point here is that the numbers in the envelopes may *not* be random. We're picking the numbers and we're choosing them in a way that we think will defeat your guessing strategy. We'll only use randomization to choose the numbers if that serves our purpose: making you lose!

Intuition Behind the Winning Strategy

People are surprised when they first learn that there is a strategy that wins more than 50% of the time, regardless of what numbers we put in the envelopes.

Suppose that you somehow knew a number x that was in between the numbers in the envelopes. Now you peek in one envelope and see a number. If it is bigger than x, then you know you’re peeking at the higher number. If it is smaller than x, then you’re peeking at the lower number. In other words, if you know a number x between the numbers in the envelopes, then you are certain to win the game.

The only flaw with this brilliant strategy is that you do *not* know such an x. This sounds like a dead end, but there’s a cool way to salvage things: try to *guess* x! There is some probability that you guess correctly. In this case, you win 100% of the time. On the other hand, if you guess incorrectly, then you’re no worse off than before; your chance of winning is still 50%. Combining these two cases, your overall chance of winning is better than 50%.

Many intuitive arguments about probability are wrong despite sounding persuasive. But this one goes the other way: it may not convince you, but it’s actually correct. To justify this, we’ll go over the argument in a more rigorous way—and while we’re at it, work out the optimal way to play.

Analysis of the Winning Strategy

For generality, suppose that we can choose numbers from the integer interval $[0..n]$. Call the lower number L and the higher number H.

Your goal is to guess a number x between L and H. It’s simplest if x does not equal L or H, so you should select x at random from among the half-integers:

$$\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots, \frac{2n-1}{2}$$

But what probability distribution should you use?

The uniform distribution—selecting each of these half-integers with equal probability—turns out to be your best bet. An informal justification is that if we figured out that you were unlikely to pick some number—say $50\frac{1}{2}$—then we’d always put 50 and 51 in the envelopes. Then you’d be unlikely to pick an x between L and H and would have less chance of winning.

After you’ve selected the number x, you peek into an envelope and see some number T. If $T > x$, then you guess that you’re looking at the larger number. If $T < x$, then you guess that the other number is larger.

All that remains is to determine the probability that this strategy succeeds. We can do this with the usual four step method and a tree diagram.

Step 1: Find the sample space.
You either choose x too low ($< L$), too high ($> H$), or just right ($L < x < H$). Then you either peek at the lower number ($T = L$) or the higher number ($T = H$). This gives a total of six possible outcomes, as show in Figure 19.3.

Step 2: Define events of interest.
The four outcomes in the event that you win are marked in the tree diagram.

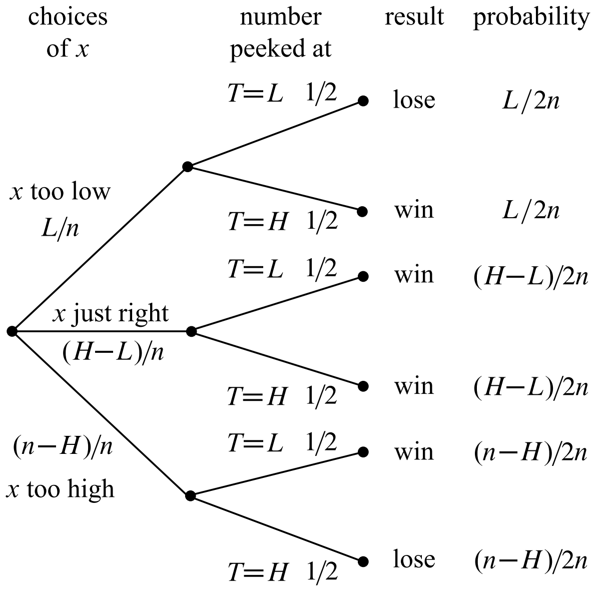


Figure 19.3 The tree diagram for the numbers game.

Step 3: Assign outcome probabilities.
First, we assign edge probabilities. Your guess x is too low with probability L/n, too high with probability $(n - H)/n$, and just right with probability $(H - L)/n$. Next, you peek at either the lower or higher number with equal probability. Multiplying along root-to-leaf paths gives the outcome probabilities.

Step 4: Compute event probabilities.
The probability of the event that you win is the sum of the probabilities of the four outcomes in that event:

$$\begin{aligned}\Pr[\text{win}] &= \frac{L}{2n} + \frac{H-L}{2n} + \frac{H-L}{2n} + \frac{n-H}{2n} \\ &= \frac{1}{2} + \frac{H-L}{2n} \\ &\geq \frac{1}{2} + \frac{1}{2n}\end{aligned}$$

The final inequality relies on the fact that the higher number H is at least 1 greater than the lower number L since they are required to be distinct.

Sure enough, you win with this strategy more than half the time, regardless of the numbers in the envelopes! So with numbers chosen from the range $0, 1, \ldots, 100$, you win with probability at least $1/2 + 1/200 = 50.5\%$. If instead we agree to stick to numbers $0, \dots, 10$, then your probability of winning rises to 55%. By Las Vegas standards, those are great odds.

Randomized Algorithms

The best strategy to win the numbers game is an example of a *randomized algorithm*—it uses random numbers to influence decisions. Protocols and algorithms that make use of random numbers are very important in computer science. There are many problems for which the best known solutions are based on a random number generator.

For example, the most commonly-used protocol for deciding when to send a broadcast on a shared bus or Ethernet is a randomized algorithm known as *exponential backoff*. One of the most commonly-used sorting algorithms used in practice, called *quicksort*, uses random numbers. You'll see many more examples if you take an algorithms course. In each case, randomness is used to improve the probability that the algorithm runs quickly or otherwise performs well.

19.3.4 Binomial Distributions

The third commonly-used distribution in computer science is the *binomial distribution*. The standard example of a random variable with a binomial distribution is the number of heads that come up in n independent flips of a coin. If the coin is fair, then the number of heads has an *unbiased binomial distribution*, specified by the pdf $f_n : [0..n] \to [0, 1]$:

$$f_n(k) ::= \binom{n}{k} 2^{-n}.$$

This is because there are $\binom{n}{k}$ sequences of n coin tosses with exactly k heads, and each such sequence has probability 2^{-n}.

A plot of $f_{20}(k)$ is shown in Figure 19.4. The most likely outcome is $k = 10$ heads, and the probability falls off rapidly for larger and smaller values of k. The falloff regions to the left and right of the main hump are called the *tails of the distribution*.

In many fields, including Computer Science, probability analyses come down to getting small bounds on the tails of the binomial distribution. In the context of a problem, this typically means that there is very small probability that something *bad* happens, which could be a server or communication link overloading or a randomized algorithm running for an exceptionally long time or producing the wrong result.

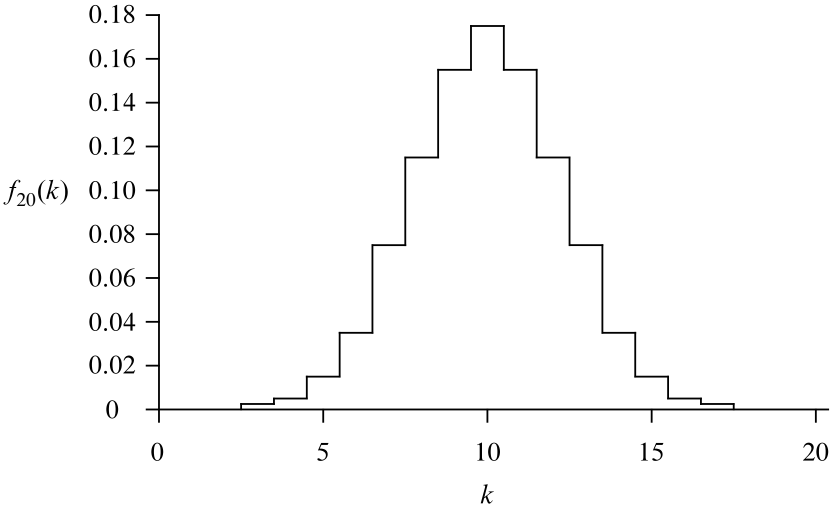


Figure 19.4 The pdf for the unbiased binomial distribution for $n = 20$, $f_{20}(k)$.

The tails do get small very fast. For example, the probability of flipping at most 25 heads in 100 tosses is less than 1 in 3,000,000. In fact, the tail of the distribution falls off so rapidly that the probability of flipping exactly 25 heads is nearly twice the probability of flipping exactly 24 heads *plus* the probability of flipping exactly 23 heads *plus* ... the probability of flipping no heads.

The General Binomial Distribution

If the coins are biased so that each coin is heads with probability p and tails with probability $q ::= 1 - p$, then the number of heads has a *general binomial density function* specified by the pdf $f_{n,p} : [0..n] \to [0,1]$ where

$$f_{n,p}(k) = \binom{n}{k} p^k q^{n-k}. \tag{19.1}$$

for some $n \in \mathbb{N}^+$ and $p \in [0,1]$. This is because there are $\binom{n}{k}$ sequences with k heads and $n - k$ tails, but now $p^k q^{n-k}$ is the probability of each such sequence.

For example, the plot in Figure 19.5 shows the probability density function $f_{n,p}(k)$ corresponding to flipping $n = 20$ independent coins that are heads with probability $p = 0.75$. The graph shows that we are most likely to get $k = 15$ heads, as you might expect. Once again, the probability falls off quickly for larger and smaller values of k.

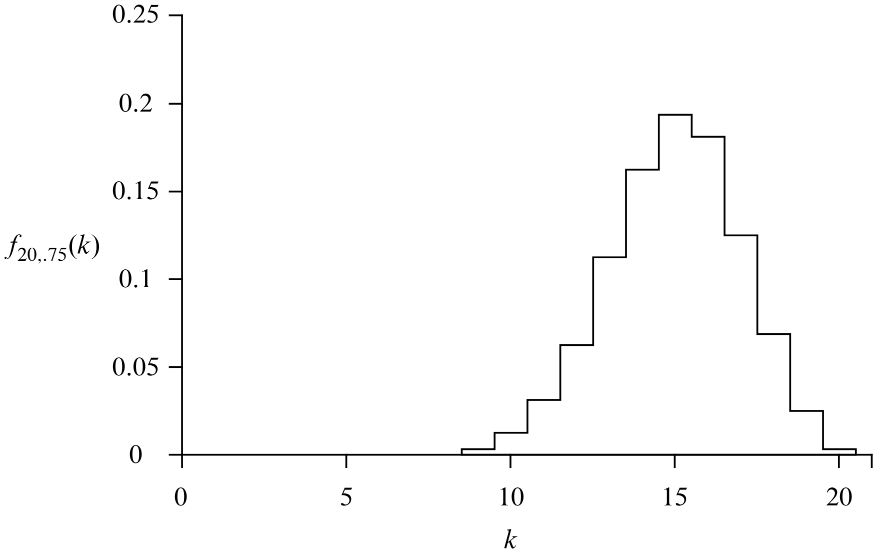


Figure 19.5 The pdf for the general binomial distribution $f_{n,p}(k)$ for $n = 20$ and $p = .75$.

19.4 Great Expectations

The *expectation* or *expected value* of a random variable in simple cases is just an average value. For example, the first thing you typically want to know when you see your grade on an exam is the average score of the class. This average score is the same as the expected score of a random student.

In general, the expected value of a random variable is the sum of all it possible values when each value is weighted according to its probability. To make this work, we need to be able to add values and multiply them by probabilities. This will certainly be possible if the values are real numbers; for technical reasons, we focus on *nonnegative* real values. Now we can define expected value formally as follows:

Definition 19.4.1. If R is a nonnegative real-valued random variable defined on a sample space $\mathcal{S}$, then the expectation of R is

$$\text{Ex}[R] ::= \sum_{\omega \in \mathcal{S}} R(\omega) \Pr[\omega]. \tag{19.2}$$

The expectation of a random variable is also known as its *mean*.

From now on, we will assume our *random variables are nonnegative real-valued* unless we explcitly say otherwise.

Let’s work through some examples.

19.4.1 The Expected Value of a Uniform Random Variable

Rolling a 6-sided die provides an example of a uniform random variable. Let R be the value that comes up when you roll a fair 6-sided die. Then by (19.2), the expected value of R is

$$\text{Ex}[R] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}.$$

This calculation shows that the name “expected” value is a little misleading; the random variable might *never* actually take on that value. No one expects to roll a $3\frac{1}{2}$ on an ordinary die!

In general, if R_n is a random variable with a uniform distribution on $\{a_1, a_2, \ldots, a_n\}$, then the expectation of R_n is simply the average of the a_i’s:

$$\text{Ex}[R_n] = \frac{a_1 + a_2 + \cdots + a_n}{n}.$$

19.4.2 The Expected Value of a Reciprocal Random Variable

Define a random variable S to be the reciprocal of the value that comes up when you roll a fair 6-sided die. That is, $S = 1/R$ where R is the value that you roll. Now,

$$\text{Ex}[S] = \text{Ex}\left[\frac{1}{R}\right] = \frac{1}{1} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{4} \cdot \frac{1}{6} + \frac{1}{5} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{49}{120}.$$

Notice that

$$\text{Ex}\left[1/R\right] \neq 1/\text{Ex}[R].$$

Assuming that these two quantities are equal is a common mistake.

19.4.3 The Expected Value of an Indicator Random Variable

The expected value of an indicator random variable for an event is just the probability of that event.

Lemma 19.4.2. *If I_A is the indicator random variable for event A, then*

$$\text{Ex}[I_A] = \Pr[A].$$

Proof.

$$\begin{aligned} \text{Ex}[I_A] = 1 \cdot \Pr[I_A = 1] + 0 \cdot \Pr[I_A = 0] &= \Pr[I_A = 1] \\ &= \Pr[A]. \qquad \text{(def of } I_A\text{)} \end{aligned}$$

For example, if A is the event that a coin with bias p comes up heads, then $\text{Ex}[I_A] = \Pr[I_A = 1] = p$.

19.4.4 Alternate Definition of Expectation

There is another standard way to define expectation:

Theorem 19.4.3. *For any random variable R,*

$$\text{Ex}[R] = \sum_{x \in \text{range}(R)} x \cdot \Pr[R = x]. \tag{19.3}$$

The proof of Theorem 19.4.3, like many of the elementary proofs about expectation in this chapter, follows by regrouping of terms in equation (19.2):

Proof. Suppose R is defined on a sample space $\mathcal{S}$. Then,

$$\begin{aligned} \text{Ex}[R] &::= \sum_{\omega \in \mathcal{S}} R(\omega) \Pr[\omega] \\ &= \sum_{x \in \text{range}(R)} \ \sum_{\omega \in [R=x]} R(\omega) \Pr[\omega] \\ &= \sum_{x \in \text{range}(R)} \ \sum_{\omega \in [R=x]} x \Pr[\omega] && \text{(def of the event } [R = x]\text{)} \\ &= \sum_{x \in \text{range}(R)} x \left(\sum_{\omega \in [R=x]} \Pr[\omega] \right) && \text{(factoring } x \text{ from the inner sum)} \\ &= \sum_{x \in \text{range}(R)} x \cdot \Pr[R = x]. && \text{(def of } \Pr[R = x]\text{)} \end{aligned}$$

The first equality follows because the events $[R = x]$ for $x \in \text{range}(R)$ partition the sample space $\mathcal{S}$, so summing over the outcomes in $[R = x]$ for $x \in \text{range}(R)$ is the same as summing over $\mathcal{S}$. ■

In general, equation (19.3) is more useful than the defining equation (19.2) for calculating expected values. It also has the advantage that it does not depend on the sample space, but only on the density function of the random variable. On the other hand, summing over all outcomes as in equation (19.2) allows easier proofs of some basic properties of expectation.

Notice that the order in which terms appear in the sums (19.3) and (19.2) is not specified, and the proof of Theorem 19.4.3—and lots of proofs below—involve regrouping the terms in sums. This is OK because of a well-known property of countable sums of nonnegative real numbers:

Theorem 19.4.4. *A countable sum of nonnegative real numbers converges to the same value, or else always diverges, regardless of the order in which the numbers are summed.*

In fact as long as reordering terms in the infinite sum (19.2) for expectation preserves convergence, we can allow random variables R taking negative as well as positive values. In this case, $\text{Ex}[R]$ will be well-defined and will have all the basic properties we establish below for nonnegative variables. But reordering does not preserve convergence for arbitrary sums of positive and negative values (see Problems 14.14 and 14.16), and there is no useful definition of the expectation for *arbitrary* real-valued random variables.

19.4.5 Conditional Expectation

Just like event probabilities, expectations can be conditioned on some event. Given a random variable R, the expected value of R conditioned on an event A is the probability-weighted average value of R over outcomes in A. More formally:

Definition 19.4.5. The *conditional expectation* $\text{Ex}[R \mid A]$ of a random variable R given event A is:

$$\text{Ex}[R \mid A] ::= \sum_{r \in \text{range}(R)} r \cdot \Pr\left[R = r \mid A\right]. \tag{19.4}$$

For example, we can compute the expected value of a roll of a fair die, given that the number rolled is at least 4. We do this by letting R be the outcome of a roll of the die. Then by equation (19.4),

$$\text{Ex}[R \mid R \geq 4] = \sum_{i=1}^{6} i \cdot \Pr\left[R = i \mid R \geq 4\right] = 1 \cdot 0 + 2 \cdot 0 + 3 \cdot 0 + 4 \cdot \tfrac{1}{3} + 5 \cdot \tfrac{1}{3} + 6 \cdot \tfrac{1}{3} = 5.$$

Conditional expectation is useful in dividing complicated expectation calculations into simpler cases. We can find a desired expectation by calculating the conditional expectation in each simple case and averaging them, weighing each case by its probability.

For example, suppose that 49.6% of the people in the world are male and the rest female—which is more or less true. Also suppose the expected height of a randomly chosen male is $5'\,11''$, while the expected height of a randomly chosen female is $5'\,5.''$ What is the expected height of a randomly chosen person? We can calculate this by averaging the heights of men and women. Namely, let H be the height (in feet) of a randomly chosen person, and let M be the event that the person is male and F the event that the person is female. Then

$$\begin{aligned}
\text{Ex}[H] &= \text{Ex}[H \mid M]\Pr[M] + \text{Ex}[H \mid F]\Pr[F] \\
&= (5 + 11/12) \cdot 0.496 + (5 + 5/12) \cdot (1 - 0.496) \\
&= 5.6646\ldots.
\end{aligned}$$

which is a little less than 5' 8."

This method is justified by:

Theorem 19.4.6 (Law of Total Expectation)**.** *Let R be a random variable on a sample space $\mathcal{S}$, and suppose that $A_1, A_2, \ldots,$ is a partition of $\mathcal{S}$. Then*

$$\text{Ex}[R] = \sum_i \text{Ex}[R \mid A_i]\Pr[A_i].$$

Proof.

$$\begin{aligned}
\text{Ex}[R] &= \sum_{r \in \text{range}(R)} r \cdot \Pr[R = r] && \text{(by 19.3)} \\
&= \sum_r r \cdot \sum_i \Pr\left[R = r \mid A_i\right]\Pr[A_i] && \text{(Law of Total Probability)} \\
&= \sum_r \sum_i r \cdot \Pr\left[R = r \mid A_i\right]\Pr[A_i] && \text{(distribute constant } r\text{)} \\
&= \sum_i \sum_r r \cdot \Pr\left[R = r \mid A_i\right]\Pr[A_i] && \text{(exchange order of summation)} \\
&= \sum_i \Pr[A_i] \sum_r r \cdot \Pr\left[R = r \mid A_i\right] && \text{(factor constant } \Pr[A_i]\text{)} \\
&= \sum_i \Pr[A_i]\,\text{Ex}[R \mid A_i]. && \text{(Def 19.4.5 of cond. expectation)}
\end{aligned}$$

■

19.4.6 Geometric Distributions

A computer program crashes at the end of each hour of use with probability p, if it has not crashed already. What is the expected time until the program crashes?

This will be easy to figure out using the Law of Total Expectation, Theorem 19.4.6. Specifically, we want to find $\text{Ex}[C]$ where C is the number of hours until the first crash. We'll do this by conditioning on whether or not the crash occurs in the first hour.

So define A to be the event that the system fails on the first step and $\overline{A}$ to be the complementary event that the system does not fail on the first step. Then the mean time to failure $\text{Ex}[C]$ is

$$\text{Ex}[C] = \text{Ex}[C \mid A]\,\Pr[A] + \text{Ex}[C \mid \overline{A}]\,\Pr[\overline{A}]. \tag{19.5}$$

Since A is the condition that the system crashes on the first step, we know that

$$\text{Ex}[C \mid A] = 1. \tag{19.6}$$

Since $\overline{A}$ is the condition that the system does *not* crash on the first step, conditioning on $\overline{A}$ is equivalent to taking a first step without failure and then starting over without conditioning. Hence,

$$\text{Ex}[C \mid \overline{A}] = 1 + \text{Ex}[C]. \tag{19.7}$$

Plugging (19.6) and (19.7) into (19.5):

$$\begin{aligned}\text{Ex}[C] &= 1 \cdot p + (1 + \text{Ex}[C])q \\ &= p + 1 - p + q\,\text{Ex}[C] \\ &= 1 + q\,\text{Ex}[C].\end{aligned}$$

Then, rearranging terms gives

$$1 = \text{Ex}[C] - q\,\text{Ex}[C] = p\,\text{Ex}[C],$$

and thus

$$\text{Ex}[C] = 1/p.$$

The general principle here is well-worth remembering.

Mean Time to Failure

If a system independently fails at each time step with probability p, then the expected number of steps up to the first failure is $1/p$.

So, for example, if there is a 1% chance that the program crashes at the end of each hour, then the expected time until the program crashes is $1/0.01 = 100$ hours.

As a further example, suppose a couple insists on having children until they get a boy, then how many baby girls should they expect before their first boy? Assume for simplicity that there is a 50% chance that a child will be a boy and that the genders of siblings are mutually independent.

This is really a variant of the previous problem. The question, “How many hours until the program crashes?” is mathematically the same as the question, “How many children must the couple have until they get a boy?” In this case, a crash corresponds to having a boy, so we should set $p = 1/2$. By the preceding analysis, the couple should expect a baby boy after having $1/p = 2$ children. Since the last of these will be a boy, they should expect just one girl. So even in societies where couples pursue this commitment to boys, the expected population will divide evenly between boys and girls.

There is a simple intuitive argument that explains the mean time to failure formula (19.8). Suppose the system is restarted after each failure. This makes the mean time to failure the same as the mean time between successive repeated failures. Now if the probability of failure at a given step is p, then after n steps we expect to have pn failures. Now the average number of steps between failures is, by definition, equal to $n/pn = 1/p$.

For the record, we’ll state a formal version of this result. A random variable like C that counts steps to first failure is said to have a *geometric distribution* with parameter p.

Definition 19.4.7. A random variable C has a *geometric distribution* with parameter p iff $\text{codomain}(C) = \mathbb{Z}^+$ and

$$\Pr[C = i] = q^{i-1}p.$$

Lemma 19.4.8. *If a random variable C has a geometric distribution with parameter p, then*

$$\text{Ex}[C] = \frac{1}{p}. \tag{19.8}$$

19.4.7 Expected Returns in Gambling Games

Some of the most interesting examples of expectation can be explained in terms of gambling games. For straightforward games where you win w dollars with probability p and you lose x dollars with probability $q = 1 - p$, it is easy to compute your *expected return* or *winnings*. It is simply

$$pw - qx \text{ dollars.}$$

For example, if you are flipping a fair coin and you win \$1 for heads and you lose \$1 for tails, then your expected winnings are

$$\frac{1}{2} \cdot 1 - \left(1 - \frac{1}{2}\right) \cdot 1 = 0.$$

In such cases, the game is said to be *fair* since your expected return is zero.

Splitting the Pot

We’ll now look at a different game which is fair—but only on first analysis.

It’s late on a Friday night in your neighborhood hangout when two new biker dudes, Eric and Nick, stroll over and propose a simple wager. Each player will put \$2 on the bar and secretly write “heads” or “tails” on their napkin. Then you will flip a fair coin. The \$6 on the bar will then be “split”—that is, be divided equally—among the players who correctly predicted the outcome of the coin toss. Pot splitting like this is a familiar feature in poker games, betting pools, and lotteries.

This sounds like a fair game, but after your regrettable encounter with strange dice (Section 17.3), you are definitely skeptical about gambling with bikers. So before agreeing to play, you go through the four-step method and write out the tree diagram to compute your expected return. The tree diagram is shown in Figure 19.6.

The “payoff” values in Figure 19.6 are computed by dividing the \$6 pot[1] among those players who guessed correctly and then subtracting the \$2 that you put into the pot at the beginning. For example, if all three players guessed correctly, then your payoff is \$0, since you just get back your \$2 wager. If you and Nick guess correctly and Eric guessed wrong, then your payoff is

$$\frac{6}{2} - 2 = 1.$$

In the case that everyone is wrong, you all agree to split the pot and so, again, your payoff is zero.

To compute your expected return, you use equation (19.3):

$$\begin{aligned}\text{Ex[payoff]} &= 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 4 \cdot \frac{1}{8} \\ &\qquad + (-2) \cdot \frac{1}{8} + (-2) \cdot \frac{1}{8} + (-2) \cdot \frac{1}{8} + 0 \cdot \frac{1}{8} \\ &= 0.\end{aligned}$$

[1] The money invested in a wager is commonly referred to as the *pot*.

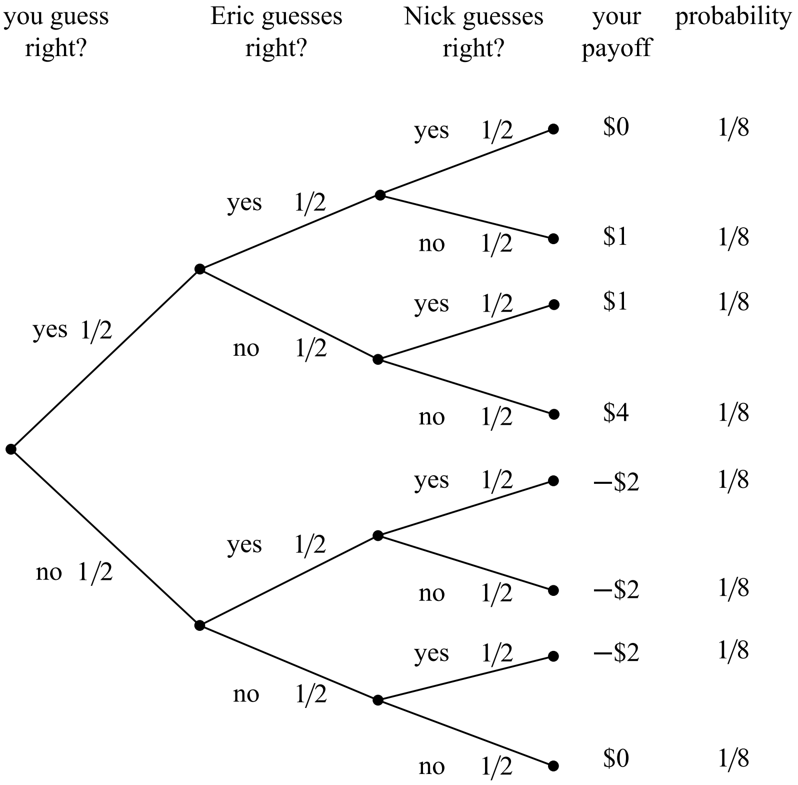


Figure 19.6 The tree diagram for the game where three players each wager $2 and then guess the outcome of a fair coin toss. The winners split the pot.

This confirms that the game is fair. So, for old time’s sake, you break your solemn vow to never ever engage in strange gambling games.

The Impact of Collusion

Needless to say, things are not turning out well for you. The more times you play the game, the more money you seem to be losing. After 1000 wagers, you have lost over \$500. As Nick and Eric are consoling you on your “bad luck,” you do a back-of-the-envelope calculation and decide that the probability of losing \$500 in 1000 fair \$2 wagers is very, very small.

Now it is possible of course that you are very, very unlucky. But it is more likely that something fishy is going on. Somehow the tree diagram in Figure 19.6 is not a good model of the game.

The “something” that’s fishy is the opportunity that Nick and Eric have to collude against you. The fact that the coin flip is fair certainly means that each of Nick and Eric can only guess the outcome of the coin toss with probability $1/2$. But when you look back at the previous 1000 bets, you notice that Eric and Nick never made the same guess. In other words, Nick always guessed “tails” when Eric guessed “heads,” and vice-versa. Modelling this fact now results in a slightly different tree diagram, as shown in Figure 19.7.

The payoffs for each outcome are the same in Figures 19.6 and 19.7, but the probabilities of the outcomes are different. For example, it is no longer possible for all three players to guess correctly, since Nick and Eric are always guessing differently. More importantly, the outcome where your payoff is \$4 is also no longer possible. Since Nick and Eric are always guessing differently, one of them will always get a share of the pot. As you might imagine, this is not good for you!

When we use equation (19.3) to compute your expected return in the collusion scenario, we find that

$$\begin{aligned}\text{Ex}[\text{payoff}] &= 0\cdot 0 + 1\cdot\frac{1}{4} + 1\cdot\frac{1}{4} + 4\cdot 0\\ &\qquad + (-2)\cdot 0 + (-2)\cdot\frac{1}{4} + (-2)\cdot\frac{1}{4} + 0\cdot 0\\ &= -\frac{1}{2}.\end{aligned}$$

So watch out for these biker dudes! By colluding, Nick and Eric have made it so that you expect to lose \$.50 every time you play. No wonder you lost \$500 over the course of 1000 wagers.

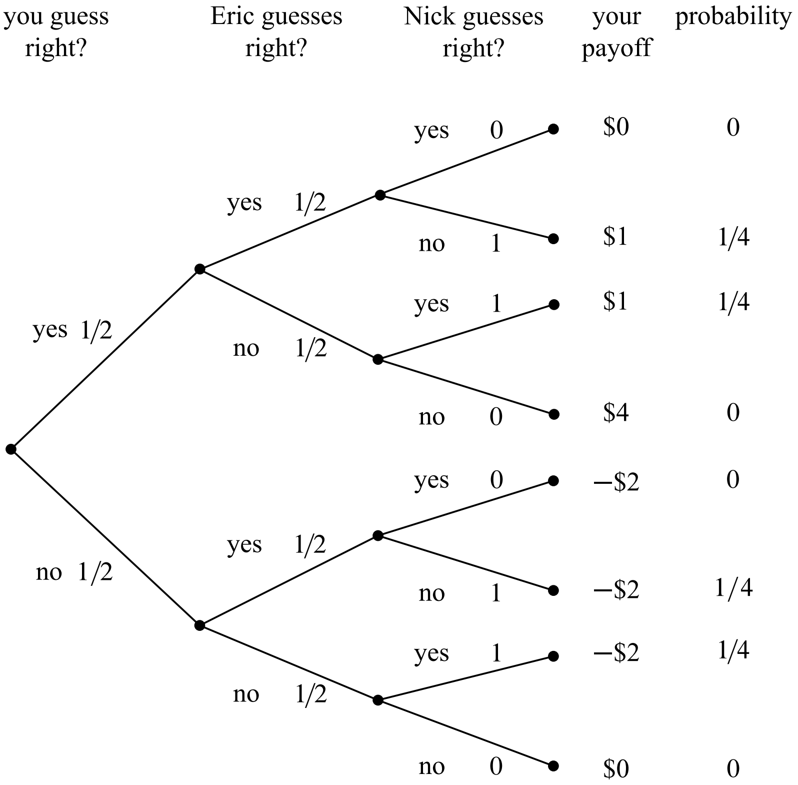


Figure 19.7 The revised tree diagram reflecting the scenario where Nick always guesses the opposite of Eric.

How to Win the Lottery

Similar opportunities to collude arise in many betting games. For example, consider the typical weekly football betting pool, where each participant wagers \$10 and the participants that pick the most games correctly split a large pot. The pool seems fair if you think of it as in Figure 19.6. But, in fact, if two or more players collude by guessing differently, they can get an “unfair” advantage at your expense!

In some cases, the collusion is inadvertent and you can profit from it. For example, many years ago, a former MIT Professor of Mathematics named Herman Chernoff figured out a way to make money by playing the state lottery. This was surprising since the state usually takes a large share of the wagers before paying the winners, and so the expected return from a lottery ticket is typically pretty poor. So how did Chernoff find a way to make money? It turned out to be easy!

In a typical state lottery,

- all players pay \$1 to play and select 4 numbers from 1 to 36,
- the state draws 4 numbers from 1 to 36 uniformly at random,
- the states divides 1/2 of the money collected among the people who guessed correctly and spends the other half redecorating the governor’s residence.

This is a lot like the game you played with Nick and Eric, except that there are more players and more choices. Chernoff discovered that a small set of numbers was selected by a large fraction of the population. Apparently many people think the same way; they pick the same numbers not on purpose as in the previous game with Nick and Eric, but based on the Red Sox winning average or today’s date. The result is as though the players were intentionally colluding to lose. If any one of them guessed correctly, then they’d have to split the pot with many other players. By selecting numbers uniformly at random, Chernoff was unlikely to get one of these favored sequences. So if he won, he’d likely get the whole pot! By analyzing actual state lottery data, he determined that he could win an average of 7 cents on the dollar. In other words, his expected return was not $-\$.50$ as you might think, but $+\$.07$.[2] Inadvertent collusion often arises in betting pools and is a phenomenon that you can take advantage of.

[2] Most lotteries now offer randomized tickets to help smooth out the distribution of selected sequences.

19.5 Linearity of Expectation

Expected values obey a simple, very helpful rule called *Linearity of Expectation*. Its simplest form says that the expected value of a sum of random variables is the sum of the expected values of the variables.

Theorem 19.5.1. *For any random variables R_1 and R_2,*

$$\text{Ex}[R_1 + R_2] = \text{Ex}[R_1] + \text{Ex}[R_2].$$

Proof. Let $T ::= R_1 + R_2$. The proof follows straightforwardly by rearranging terms in equation (19.2) in the definition of expectation:

$$\begin{aligned} \text{Ex}[T] &::= \sum_{\omega \in \mathcal{S}} T(\omega) \cdot \Pr[\omega] \\ &= \sum_{\omega \in \mathcal{S}} (R_1(\omega) + R_2(\omega)) \cdot \Pr[\omega] && \text{(def of } T\text{)} \\ &= \sum_{\omega \in \mathcal{S}} R_1(\omega) \Pr[\omega] + \sum_{\omega \in \mathcal{S}} R_2(\omega) \Pr[\omega] && \text{(rearranging terms)} \\ &= \text{Ex}[R_1] + \text{Ex}[R_2]. && \text{(by (19.2))} \end{aligned}$$

■

Essentially the same proof implies:

Theorem 19.5.2. *For random variables R_1, R_2 and constants $a_1, a_2 \in \mathbb{R}$,*

$$\text{Ex}[a_1 R_1 + a_2 R_2] = a_1 \text{Ex}[R_1] + a_2 \text{Ex}[R_2].$$

In other words, expectation is a linear function. A routine induction extends the result to more than two variables:

Corollary 19.5.3 (Linearity of Expectation)**.** *For any random variables $R_1, \ldots, R_k$ and constants $a_1, \ldots, a_k \in \mathbb{R}$,*

$$\text{Ex}\left[\sum_{i=1}^{k} a_i R_i\right] = \sum_{i=1}^{k} a_i \text{Ex}[R_i].$$

The great thing about linearity of expectation is that *no independence is required*. This is really useful, because dealing with independence is a pain, and we often need to work with random variables that are not independent.

As an example, let's compute the expected value of the sum of two fair dice.

19.5.1 Expected Value of Two Dice

What is the expected value of the sum of two fair dice?

Let the random variable R_1 be the number on the first die, and let R_2 be the number on the second die. We observed earlier that the expected value of one die is 3.5. We can find the expected value of the sum using linearity of expectation:

$$\text{Ex}[R_1 + R_2] = \text{Ex}[R_1] + \text{Ex}[R_2] = 3.5 + 3.5 = 7.$$

Assuming that the dice were independent, we could use a tree diagram to prove that this expected sum is seven, but this would be a bother since there are 36 cases. And without assuming independence, it's not apparent how to apply the tree diagram approach at all. But notice that we did *not* have to assume that the two dice were independent. For example, suppose the roll of the second die was forced to match the roll of the first die. Then the expected sum of two dice remains equal to seven because the second die is still fair.

19.5.2 Sums of Indicator Random Variables

Linearity of expectation is especially useful when you have a sum of indicator random variables. As an example, suppose there is a dinner party where n men check their hats. The hats are mixed up during dinner, so that afterward each man receives a random hat. In particular, each man gets his own hat with probability $1/n$. What is the expected number of men who get their own hat?

Letting G be the number of men that get their own hat, we want to find the expectation of G. But all we know about G is that the probability that a man gets his own hat back is $1/n$. There are many different probability distributions of hat permutations with this property, so we don't know enough about the distribution of G to calculate its expectation directly using equation (19.2) or (19.3). But linearity of expectation lets us sidestep this issue.

We'll use a standard, useful trick to apply linearity, namely, we'll express G as a sum of indicator variables. In particular, let G_i be an indicator for the event that the ith man gets his own hat. That is, $G_i = 1$ if the ith man gets his own hat, and $G_i = 0$ otherwise. The number of men that get their own hat is then the sum of these indicator random variables:

$$G = G_1 + G_2 + \cdots + G_n. \tag{19.9}$$

These indicator variables are *not* mutually independent. For example, if $n - 1$ men all get their own hats, then the last man is certain to receive his own hat. But again, we don't need to worry about this dependence, since linearity holds regardless.

Since G_i is an indicator random variable, we know from Lemma 19.4.2 that

$$\text{Ex}[G_i] = \Pr[G_i = 1] = 1/n. \tag{19.10}$$

By Linearity of Expectation and equation (19.9), this means that

$$\begin{aligned}
\text{Ex}[G] &= \text{Ex}[G_1 + G_2 + \cdots + G_n] \\
&= \text{Ex}[G_1] + \text{Ex}[G_2] + \cdots + \text{Ex}[G_n] \\
&= \overbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}^{n} \\
&= 1.
\end{aligned}$$

So even though we don't know much about how hats are scrambled, we've figured out that on average, just one man gets his own hat back, regardless of the number of men with hats!

More generally, Linearity of Expectation provides a very good method for computing the expected number of events that will happen.

Theorem 19.5.4. *Given any collection of events $A_1, A_2, \ldots, A_n$, the expected number of events that will occur is*

$$\sum_{i=1}^{n} \Pr[A_i].$$

For example, A_i could be the event that the ith man gets the right hat back. But in general, it could be any subset of the sample space, and we are asking for the expected number of events that will contain a random sample point.

Proof. Define R_i to be the indicator random variable for A_i, where $R_i(\omega) = 1$ if $w \in A_i$ and $R_i(\omega) = 0$ if $w \notin A_i$. Let $R = R_1 + R_2 + \cdots + R_n$. Then

$$\begin{aligned}
\text{Ex}[R] &= \sum_{i=1}^{n} \text{Ex}[R_i] && \text{(by Linearity of Expectation)} \\
&= \sum_{i=1}^{n} \Pr[R_i = 1] && \text{(by Lemma 19.4.2)} \\
&= \sum_{i=1}^{n} \Pr[A_i]. && \text{(def of indicator variable)}
\end{aligned}$$

So whenever you are asked for the expected number of events that occur, all you have to do is sum the probabilities that each event occurs. Independence is not needed.

19.5.3 Expectation of a Binomial Distribution

Suppose that we independently flip n biased coins, each with probability p of coming up heads. What is the expected number of heads?

Let J be the random variable denoting the number of heads. Then J has a binomial distribution with parameters n, p, and

$$\Pr[J = k] = \binom{n}{k} p^k q^{n-k}.$$

Applying equation (19.3), this means that

$$\text{Ex}[J] = \sum_{k=0}^{n} k \Pr[J = k] = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k}. \tag{19.11}$$

This sum looks a tad nasty, but linearity of expectation leads to an easy derivation of a simple closed form. We just express J as a sum of indicator random variables, which is easy. Namely, let J_i be the indicator random variable for the ith coin coming up heads, that is,

$$J_i ::= \begin{cases} 1 & \text{if the } i\text{th coin is heads} \\ 0 & \text{if the } i\text{th coin is tails.} \end{cases}$$

Then the number of heads is simply

$$J = J_1 + J_2 + \cdots + J_n.$$

By Theorem 19.5.4,

$$\text{Ex}[J] = \sum_{i=1}^{n} \Pr[J_i] = pn. \tag{19.12}$$

That really was easy. If we flip n mutually independent coins, we expect to get pn heads. Hence the expected value of a binomial distribution with parameters n and p is simply pn.

But what if the coins are not mutually independent? It doesn't matter—the answer is still pn because Linearity of Expectation and Theorem 19.5.4 do not assume any independence.

If you are not yet convinced that Linearity of Expectation and Theorem 19.5.4 are powerful tools, consider this: without even trying, we have used them to prove a complicated looking identity, namely,

$$\sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = pn, \tag{19.13}$$

which follows by combining equations (19.11) and (19.12) (see also Exercise 19.31).

The next section has an even more convincing illustration of the power of linearity to solve a challenging problem.

19.5.4 The Coupon Collector Problem

Every time we purchase a kid’s meal at Taco Bell, we are graciously presented with a miniature “Racin’ Rocket” car together with a launching device which enables us to project our new vehicle across any tabletop or smooth floor at high velocity. Truly, our delight knows no bounds.

There are different colored Racin’ Rocket cars. The color of car awarded to us by the kind server at the Taco Bell register appears to be selected uniformly and independently at random. What is the expected number of kid’s meals that we must purchase in order to acquire at least one of each color of Racin’ Rocket car?

The same mathematical question shows up in many guises: for example, what is the expected number of people you must poll in order to find at least one person with each possible birthday? The general question is commonly called the *coupon collector problem* after yet another interpretation.

A clever application of linearity of expectation leads to a simple solution to the coupon collector problem. Suppose there are five different colors of Racin’ Rocket cars, and we receive this sequence:

$$\text{blue} \quad \text{green} \quad \text{green} \quad \text{red} \quad \text{blue} \quad \text{orange} \quad \text{blue} \quad \text{orange} \quad \text{gray}.$$

Let’s partition the sequence into 5 segments:

$$\underbrace{\text{blue}}_{X_0} \quad \underbrace{\text{green}}_{X_1} \quad \underbrace{\text{green} \quad \text{red}}_{X_2} \quad \underbrace{\text{blue} \quad \text{orange}}_{X_3} \quad \underbrace{\text{blue} \quad \text{orange} \quad \text{gray}}_{X_4}.$$

The rule is that a segment ends whenever we get a new kind of car. For example, the middle segment ends when we get a red car for the first time. In this way, we can break the problem of collecting every type of car into stages. Then we can analyze each stage individually and assemble the results using linearity of expectation.

In the general case there are n colors of Racin’ Rockets that we’re collecting. Let X_k be the length of the kth segment. The total number of kid’s meals we must purchase to get all n Racin’ Rockets is the sum of the lengths of all these segments:

$$T = X_0 + X_1 + \cdots + X_{n-1}.$$

Now let’s focus our attention on X_k, the length of the kth segment. At the beginning of segment k, we have k different types of car, and the segment ends when we acquire a new type. When we own k types, each kid’s meal contains a type that we already have with probability k/n. Therefore, each meal contains a new type of car with probability $1 - k/n = (n-k)/n$. Thus, the expected number of meals until we get a new kind of car is $n/(n-k)$ by the Mean Time to Failure rule. This means that

$$\text{Ex}[X_k] = \frac{n}{n-k}.$$

Linearity of expectation, together with this observation, solves the coupon collector problem:

$$\begin{aligned}
\text{Ex}[T] &= \text{Ex}[X_0 + X_1 + \cdots + X_{n-1}] \\
&= \text{Ex}[X_0] + \text{Ex}[X_1] + \cdots + \text{Ex}[X_{n-1}] \\
&= \frac{n}{n-0} + \frac{n}{n-1} + \cdots + \frac{n}{3} + \frac{n}{2} + \frac{n}{1} \\
&= n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{3} + \frac{1}{2} + \frac{1}{1}\right) \\
&= n\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1} + \frac{1}{n}\right) \\
&= nH_n \qquad (19.14) \\
&\sim n \ln n.
\end{aligned}$$

Cool! It’s those Harmonic Numbers again.

We can use equation (19.14) to answer some concrete questions. For example, the expected number of die rolls required to see every number from 1 to 6 is:

$$6H_6 = 14.7\ldots.$$

And the expected number of people you must poll to find at least one person with each possible birthday is:

$$365H_{365} = 2364.6\ldots.$$

19.5.5 Infinite Sums

Linearity of expectation also works for an infinite number of random variables provided that the variables satisfy an absolute convergence criterion.

Theorem 19.5.5 (Linearity of Expectation)**.** *Let R_0, R_1, ..., be random variables such that*

$$\sum_{i=0}^{\infty} \operatorname{Ex}[\,|R_i|\,]$$

converges. Then

$$\operatorname{Ex}\left[\sum_{i=0}^{\infty} R_i\right] = \sum_{i=0}^{\infty} \operatorname{Ex}[R_i].$$

Proof. Let $T ::= \sum_{i=0}^{\infty} R_i$.

We leave it to the reader to verify that, under the given convergence hypothesis, all the sums in the following derivation are absolutely convergent, which justifies rearranging them as follows:

$$\begin{aligned}
\sum_{i=0}^{\infty} \operatorname{Ex}[R_i] &= \sum_{i=0}^{\infty} \sum_{s \in \mathcal{S}} R_i(s) \cdot \Pr[s] && \text{(Def. 19.4.1)} \\
&= \sum_{s \in \mathcal{S}} \sum_{i=0}^{\infty} R_i(s) \cdot \Pr[s] && \text{(exchanging order of summation)} \\
&= \sum_{s \in \mathcal{S}} \left[\sum_{i=0}^{\infty} R_i(s)\right] \cdot \Pr[s] && \text{(factoring out } \Pr[s]\text{)} \\
&= \sum_{s \in \mathcal{S}} T(s) \cdot \Pr[s] && \text{(Def. of } T\text{)} \\
&= \operatorname{Ex}[T] && \text{(Def. 19.4.1)} \\
&= \operatorname{Ex}\left[\sum_{i=0}^{\infty} R_i\right]. && \text{(Def. of } T\text{).} \quad \blacksquare
\end{aligned}$$

19.5.6 A Gambling Paradox

One of the simplest casino bets is on "red" or "black" at the roulette table. In each play at roulette, a small ball is set spinning around a roulette wheel until it lands in a red, black, or green colored slot. The payoff for a bet on red or black matches the bet; for example, if you bet $10 on red and the ball lands in a red slot, you get back your original $10 bet plus another matching $10.

The casino gets its advantage from the green slots, which make the probability of both red and black each less than 1/2. In the US, a roulette wheel has 2 green slots among 18 black and 18 red slots, so the probability of red is $18/38 \approx 0.473$. In Europe, where roulette wheels have only 1 green slot, the odds for red are a little better—that is, $18/37 \approx 0.486$—but still less than even.

Of course you can't expect to win playing roulette, even if you had the good fortune to gamble against a *fair* roulette wheel. To prove this, note that with a fair wheel, you are equally likely win or lose each bet, so your expected win on any spin is zero. Therefore if you keep betting, your expected win is the sum of your expected wins on each bet: still zero.

Even so, gamblers regularly try to develop betting strategies to win at roulette despite the bad odds. A well known strategy of this kind is *bet doubling*, where you bet, say, \$10 on red and keep doubling the bet until a red comes up. This means you stop playing if red comes up on the first spin, and you leave the casino with a \$10 profit. If red does not come up, you bet \$20 on the second spin. Now if the second spin comes up red, you get your \$20 bet plus \$20 back and again walk away with a net profit of $\$20 - 10 = \$10$. If red does not come up on the second spin, you next bet \$40 and walk away with a net win of $\$40 - 20 - 10 = \$10$ if red comes up on on the third spin, and so on.

Since we've reasoned that you can't even win against a fair wheel, this strategy against an unfair wheel shouldn't work. But wait a minute! There is a 0.486 probability of red appearing on each spin of the wheel, so the mean time until a red occurs is less than three. What's more, red will come up *eventually* with probability one, and as soon as it does, you leave the casino \$10 ahead. In other words, by bet doubling you are *certain* to win \$10, and so your expectation is \$10, not zero!

Something's wrong here.

19.5.7 Solution to the Paradox

The argument claiming the expectation is zero against a fair wheel is flawed by an implicit, invalid use of linearity of expectation for an infinite sum.

To explain this carefully, let B_n be the number of dollars you win on your nth bet, where B_n is defined to be zero if red comes up before the nth spin of the wheel. Now the dollar amount you win in any gambling session is

$$\sum_{n=1}^{\infty} B_n,$$

and your expected win is

$$\text{Ex}\left[\sum_{n=1}^{\infty} B_n\right]. \tag{19.15}$$

Moreover, since we’re assuming the wheel is fair, it’s true that $\text{Ex}[B_n] = 0$, so

$$\sum_{n=1}^{\infty} \text{Ex}[B_n] = \sum_{n=1}^{\infty} 0 = 0. \tag{19.16}$$

The flaw in the argument that you can’t win is the implicit appeal to linearity of expectation to conclude that the expectation (19.15) equals the sum of expectations in (19.16). This is a case where linearity of expectation fails to hold—even though the expectation (19.15) is 10 and the sum (19.16) of expectations converges. The problem is that the expectation of the sum of the absolute values of the bets diverges, so the condition required for infinite linearity fails. In particular, under bet doubling your nth bet is $10 \cdot 2^{n-1}$ dollars while the probability that you will make an nth bet is 2^{-n}. So

$$\text{Ex}[|B_n|] = 10 \cdot 2^{n-1} 2^{-n} = 5.$$

Therefore the sum

$$\sum_{n=1}^{\infty} \text{Ex}[|B_n|] = 5 + 5 + 5 + \cdots$$

diverges rapidly.

So the presumption that you can’t beat a fair game, and the argument we offered to support this presumption, are mistaken: by bet doubling, you can be sure to walk away a winner. Probability theory has led to an apparently absurd conclusion.

But probability theory shouldn’t be rejected because it leads to this absurd conclusion. If you only had a finite amount of money to bet with—say enough money to make k bets before going bankrupt—then it would be correct to calculate your expection by summing $B_1 + B_2 + \cdots + B_k$, and your expectation would be zero for the fair wheel and negative against an unfair wheel. In other words, in order to follow the bet doubling strategy, you need to have an infinite bankroll. So it’s absurd to assume you could actually follow a bet doubling strategy, and we needn’t be concerned when an absurd assumption leads to an absurd conclusion.

19.5.8 Expectations of Products

While the expectation of a sum is the sum of the expectations, the same is usually not true for products. For example, suppose that we roll a fair 6-sided die and denote the outcome with the random variable R. Does $\text{Ex}[R \cdot R] = \text{Ex}[R] \cdot \text{Ex}[R]$?

We know that $\text{Ex}[R] = 3\frac{1}{2}$ and thus $(\text{Ex}[R])^2 = 12\frac{1}{4}$. Let’s compute $\text{Ex}[R^2]$ to see if we get the same result.

$$\begin{aligned}\operatorname{Ex}\left[R^2\right] &= \sum_{\omega \in \mathcal{S}} R^2(\omega)\Pr[w] = \sum_{i=1}^{6} i^2 \cdot \Pr[R_i = i] \\ &= \frac{1^2}{6} + \frac{2^2}{6} + \frac{3^2}{6} + \frac{4^2}{6} + \frac{5^2}{6} + \frac{6^2}{6} = 15\ 1/6 \neq 12\ 1/4.\end{aligned}$$

That is,

$$\operatorname{Ex}[R \cdot R] \neq \operatorname{Ex}[R] \cdot \operatorname{Ex}[R].$$

So the expectation of a product is not always equal to the product of the expectations.

There is a special case when such a relationship *does* hold however; namely, when the random variables in the product are *independent*.

Theorem 19.5.6. *For any two* independent *random variables R_1, R_2,*

$$\operatorname{Ex}[R_1 \cdot R_2] = \operatorname{Ex}[R_1] \cdot \operatorname{Ex}[R_2].$$

The proof follows by rearrangement of terms in the sum that defines $\operatorname{Ex}[R_1 \cdot R_2]$. Details appear in Problem 19.29.

Theorem 19.5.6 extends routinely to a collection of mutually independent variables.

Corollary 19.5.7. *[Expectation of Independent Product]*
If random variables $R_1, R_2, \ldots, R_k$ are mutually independent, then

$$\operatorname{Ex}\left[\prod_{i=1}^{k} R_i\right] = \prod_{i=1}^{k} \operatorname{Ex}[R_i].$$

19.6 Really Great Expectations

Making independent tosses of a fair coin until some desired pattern comes up is a simple process you should feel in some command of by now, right? So how about a bet about the simplest such process—tossing until a head comes up? Ok, you're wary of betting with us, but how about this: we'll let *you set the odds.*

19.6.1 Repeating Yourself

Here's the bet: you make independent tosses of a fair coin until a head comes up. Then you will repeat the process. If a second head comes up in the same or fewer tosses than the first, you have to start over yet again. You keep starting over until you finally toss a run of tails longer than your first one. The payment rules are that you will pay me 1 cent each time you start over. When you win by finally getting a run of tails longer than your first one, I will pay you some generous amount. Notice by the way that you're certain to win—whatever your initial run of tails happened to be, a longer run will eventually occur again with probability 1!

For example, if your first tosses are `TTTH`, then you will keep tossing until you get a run of 4 tails. So your winning flips might be

`TTTHTHTTHHTTHTHTTTHTHHHTTTT.`

In this run there are 10 heads, which means you had to start over 9 times. So you would have paid me 9 cents by the time you finally won by tossing 4 tails. Now you've won, and I'll pay you generously —how does 25 cents sound? Maybe you'd rather have \$1? How about \$1000?

Of course there's a trap here. Let's calculate your expected winnings.

Suppose your initial run of tails had length k. After that, each time a head comes up, you have to start over and try to get $k+1$ tails in a row. If we regard your getting $k+1$ tails in a row as a "failed" try, and regard your having to start over because a head came up too soon as a "successful" try, then the number of times you have to start over is the number of tries till the first failure. So the expected number of tries will be the mean time to failure, which is 2^{k+1}, because the probability of tossing $k+1$ tails in a row is $2^{-(k+1)}$.

Let T be the length of your initial run of tails. So $T = k$ means that your initial tosses were $\texttt{T}^k\texttt{H}$. Let R be the number of times you repeat trying to beat your original run of tails. The number of cents you expect to finish with is the number of cents in my generous payment minus $\text{Ex}[R]$. It's now easy to calculate $\text{Ex}[R]$ by conditioning on the value of T:

$$\text{Ex}[R] = \sum_{k \in \mathbb{N}} \text{Ex}[R \mid T = k] \cdot \Pr[T = k] = \sum_{k \in \mathbb{N}} 2^{k+1} \cdot 2^{-(k+1)} = 1+1+1+\cdots = \infty.$$

So you can expect to pay me an infinite number of cents before winning my "generous" payment. No amount of generosity can make this bet fair! In fact this particular example is a special case of an astonishingly general one: the expected waiting time for *any* random variable to achieve a larger value remains infinite.

Problems for Section 19.2

Practice Problems

Problem 19.1.
Let I_A and I_B be the indicator variables for events A and B. Prove that I_A and I_B are independent iff A and B are independent.

Hint: Let $A^1 ::= A$ and $A^0 ::= \overline{A}$, so the event $[I_A = c]$ is the same as A^c for $c \in \{0, 1\}$; likewise for B^1, B^0.

Homework Problems

Problem 19.2.
Let R, S and T be random variables with the same codomain V.

(a) Suppose R is uniform—that is,

$$\Pr[R = b] = \frac{1}{|V|},$$

for all $b \in V$—and R is independent of S. Originally this text had the following argument:

> The probability that $R = S$ is the same as the probability that R takes whatever value S happens to have, therefore
>
> $$\Pr[R = S] = \frac{1}{|V|}. \qquad (19.17)$$

Are you convinced by this argument? Write out a careful proof of (19.17).

Hint: The event $[R = S]$ is a disjoint union of events

$$[R = S] = \bigcup_{b \in V} [R = b \text{ AND } S = b].$$

(b) Let $S \times T$ be the random variable giving the values of S and T.[3] Now suppose R has a uniform distribution, and R is independent of $S \times T$. How about this argument?

[3] That is, $S \times T : \mathcal{S} \to V \times V$ where

$$(S \times T)(\omega) ::= (S(\omega), T(\omega))$$

for every outcome $\omega \in \mathcal{S}$.

> The probability that $R = S$ is the same as the probability that R equals the first coordinate of whatever value $S \times T$ happens to have, and this probability remains equal to $1/|V|$ by independence. Therefore the event $[R = S]$ is independent of $[S = T]$.

Write out a careful proof that $[R = S]$ is independent of $[S = T]$.

(c) Let $V = \{1, 2, 3\}$ and (R, S, T) take the following triples of values with equal probability,

$$(1,1,1), (2,1,1), (1,2,3), (2,2,3), (1,3,2), (2,3,2).$$

Verify that

1. R is independent of $S \times T$,
2. The event $[R = S]$ is not independent of $[S = T]$.
3. S and T have a uniform distribution.

Problem 19.3.
Let R, S and T be mutually independent indicator variables.

In general, the event that $S = T$ is not independent of $R = S$. We can explain this intuitively as follows: suppose for simplicity that S is uniform, that is, equally likely to be 0 or 1. This implies that S is equally likely as not to equal R, that is $\Pr[R = S] = 1/2$; likewise, $\Pr[S = T] = 1/2$.

Now suppose further that both R and T are more likely to equal 1 than to equal 0. This implies that $R = S$ makes it more likely than not that $S = 1$, and knowing that $S = 1$, makes it more likely than not that $S = T$. So knowing that $R = S$ makes it more likely than not that $S = T$, that is, $\Pr\left[S = T \mid R = S\right] > 1/2$.

Now prove rigorously (without any appeal to intuition)

Lemma 19.6.1. *Events $[R = S]$ and $[S = T]$ are independent iff either R is uniform*[4]*, or T is uniform, or S is constant*[5]*.*

[4] That is, $\Pr[R = 1] = 1/2$.

[5] That is, $\Pr[S = 1]$ is one or zero.

Problems for Section 19.3

Practice Problems

Problem 19.4.
Suppose R, S and T be mutually independent random variables on the same probability space with uniform distribution on the range $\{1, 2, 3\}$.

Let $M = \max\{R, S, T\}$. Compute the values of the probability density function PDF_M of M.

Class Problems

> Guess the Bigger Number Game
>
> Team 1:
>
> - Write two different integers between 0 and 7 on separate pieces of paper.
> - Put the papers face down on a table.
>
> Team 2:
>
> - Turn over one paper and look at the number on it.
> - Either stick with this number or switch to the other (unseen) number.
>
> Team 2 wins if it chooses the larger number; else, Team 1 wins.

Problem 19.5.
The analysis in Section 19.3.3 implies that Team 2 has a strategy that wins 4/7 of the time no matter how Team 1 plays. Can Team 2 do better? The answer is "no," because Team 1 has a strategy that guarantees that it wins at least 3/7 of the time, no matter how Team 2 plays. Describe such a strategy for Team 1 and explain why it works.

Problem 19.6.
Suppose you have a biased coin that has probability p of flipping heads. Let J be the number of heads in n independent coin flips. So J has the general binomial

distribution:

$$\text{PDF}_J(k) = \binom{n}{k} p^k q^{n-k}$$

where $q ::= 1 - p$.

(a) Show that

$$\begin{aligned} \text{PDF}_J(k-1) &< \text{PDF}_J(k) && \text{for } k < np + p, \\ \text{PDF}_J(k-1) &> \text{PDF}_J(k) && \text{for } k > np + p. \end{aligned}$$

(b) Conclude that the maximum value of PDF_J is asymptotically equal to

$$\frac{1}{\sqrt{2\pi npq}}.$$

Hint: For the asymptotic estimate, it's ok to assume that np is an integer, so by part (a), the maximum value is $\text{PDF}_J(np)$. Use Stirling's Formula.

Problem 19.7.
Let $R_1, R_2, \ldots, R_m$, be mutually independent random variables with uniform distribution on $[1..n]$. Let $M ::= \max\{R_i \mid i \in [1..m]\}$.

(a) Write a formula for $\text{PDF}_M(1)$.

(b) More generally, write a formula for $\Pr[M \le k]$.

(c) For $k \in [1..n]$, write a formula for $\text{PDF}_M(k)$ in terms of expressions of the form "$\Pr[M \le j]$" for $j \in [1..n]$.

Homework Problems

Problem 19.8.
An over-caffeinated sailor of Tech Dinghy wanders along Seaside Boulevard. In each step, the sailor randomly moves one unit left or right with equal probability.

We let the sailor's initial position be designated location zero, with successive positions to the right labelled 1,2,..., and positions to the left labelled -1,-2,.... Let L_t be the random variable giving the sailor's location after t steps. Before he starts, the sailor is known to be at location zero, so

$$\text{PDF}_{L_0}(n) = \begin{cases} 1 & \text{if } n = 0, \\ 0 & \text{otherwise.} \end{cases}$$

After one step, the sailor is equally likely to be at location 1 or -1, so

$$\mathrm{PDF}_{L_1}(n) = \begin{cases} 1/2 & \text{if } n = \pm 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Give the distributions PDF_{L_t} for $t = 2, 3, 4$ by filling in the table of probabilities below, where omitted entries are 0. For each row, write all the nonzero entries so they have the same denominator.

	location								
	-4	-3	-2	-1	0	1	2	3	4
initially					1				
after 1 step				1/2	0	1/2			
after 2 steps			?	?	?	?	?		
after 3 steps		?	?	?	?	?	?	?	
after 4 steps	?	?	?	?	?	?	?	?	?

(b) Help the staff of the Sailing Pavilion locate the sailor by answering the following questions. Provide your derivations and reasoning.

(i) What is the final location of a t-step walk that moves right exactly i times?

(ii) How many different length-t walks are there that end at that location?

(iii) What is the probability that the sailor ends at this location?

(iv) Let $B_t ::= (L_t + t)/2$. Conclude that B_t has an unbiased binomial distribution.

Problems for Section 19.4

Practice Problems

Problem 19.9.
Bruce Lee, on a movie that didn’t go public, is practicing by breaking 5 boards with his fists. He is able to break a board with probability 0.8—he is practicing with his left fist, that’s why it’s not 1—and he breaks each board independently.

(a) What is the probability that Bruce breaks exactly 2 out of the 5 boards that are placed before him?

(b) What is the probability that Bruce breaks at most 3 out of the 5 boards that are placed before him?

(c) What is the expected number of boards Bruce will break?

Problem 19.10.
A news article reporting on the departure of a school official from California to Alabama dryly commented that this move would raise the average IQ in both states. Explain.

Class Problems

Problem 19.11.
Here’s a dice game with maximum payoff k: make three independent rolls of a fair die, and if you roll a six

- no times, then you lose 1 dollar;
- exactly once, then you win 1 dollar;
- exactly twice, then you win 2 dollars;
- all three times, then you win k dollars.

For what value of k is this game fair?[6]

Problem 19.12. (a) Suppose we flip a fair coin and let $N_{\texttt{TT}}$ be the number of flips until the first time two consecutive Tails appear. What is $\text{Ex}[N_{\texttt{TT}}]$?

Hint: Let D be the tree diagram for this process. Explain why D can be described by the tree in Figure 19.8. Use the **Law of Total Expectation** 19.4.6.

(b) Let $N_{\texttt{TH}}$ be the number of flips until a Tail immediately followed by a Head comes up. What is $\text{Ex}[N_{\texttt{TH}}]$?

(c) Suppose we now play a game: flip a fair coin until either TT or TH occurs. You win if TT comes up first, and lose if TH comes up first. Since TT takes 50% longer on average to turn up, your opponent agrees that he has the advantage. So you tell him you’re willing to play if you pay him \$5 when he wins, and he pays you with a mere 20% premium—that is \$6—when you win.

If you do this, you’re sneakily taking advantage of your opponent’s untrained intuition, since you’ve gotten him to agree to unfair odds. What is your expected profit per game?

[6] This game is actually offered in casinos with $k = 3$, where it is called Carnival Dice.

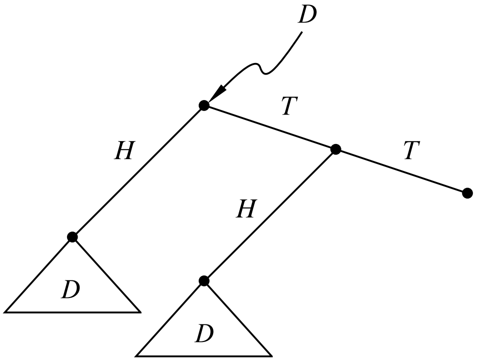


Figure 19.8 Sample space tree for coin toss until two consecutive tails.

Problem 19.13.
Ben Bitdiddle is asked to analyze a game in which a fair coin is tossed until the first time a head turns up. If this head occurs on the nth toss, and n is odd, then he wins $\$2^n/n$, but if n is even, he loses $\$2^n/n$. Ben observes that the expected dollar win from this game is

$$(1/2)\cdot 2-(1/4)\cdot 2+(1/8)\cdot 8/3+\cdots\pm(1/2^n)\cdot 2^n/n = 1-1/2+1/3-1/4+\cdots\pm 1/n.$$

which is the alternating harmonic series—a series that converges to a definite real number $r > 0$. Since $r > 0$, Ben concludes that it's to his advantage to play this game, but as usual, his shoot-from-the-hip analysis is off the mark. Explain.

Problem 19.14.
Let T be a positive integer valued random variable such that

$$\text{PDF}_T(n) = \frac{1}{an^2},$$

where

$$a ::= \sum_{n\in\mathbb{Z}^+} \frac{1}{n^2}.$$

(a) Prove that $\text{Ex}[T]$ is infinite.

(b) Prove that $\text{Ex}[\sqrt{T}]$ is finite.

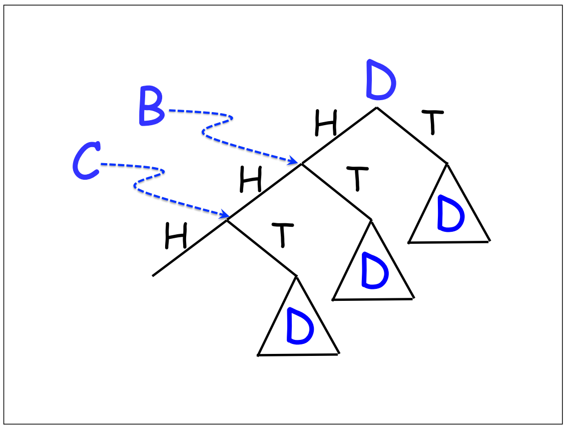


Figure 19.9 Outcome Tree for Flipping Until HHH

Exam Problems

Problem 19.15.
A record of who beat whom in a round-robin tournament can be described with a *tournament digraph*, where the vertices correspond to players and there is an edge $\langle x \rightarrow y \rangle$ iff x beat y in their game. A *ranking* of the players is a path that includes all the players. A tournament digraph may in general have one or more rankings.[7]

Suppose we construct a random tournament digraph by letting each of the players in a match be equally likely to win and having results of all the matches be mutually independent. Find a formula for the expected number of rankings in a random 10-player tournament. Conclude that there is a 10-vertex tournament digraph with more than 7000 rankings.

This problem is an instance of the *probabilistic method*. It uses probability to prove the existence of an object without constructing it.

Problem 19.16.
A coin with probability p of flipping Heads and probability $q ::= 1 - p$ of flipping tails is repeatedly flipped until three consecutive Heads occur. The outcome tree D for this setup is illustrated in Figure 19.9.

Let $e(S)$ be the expected number of flips starting at the root of subtree S of D.

[7] It has a unique ranking iff it is a DAG, see Problem 10.10.

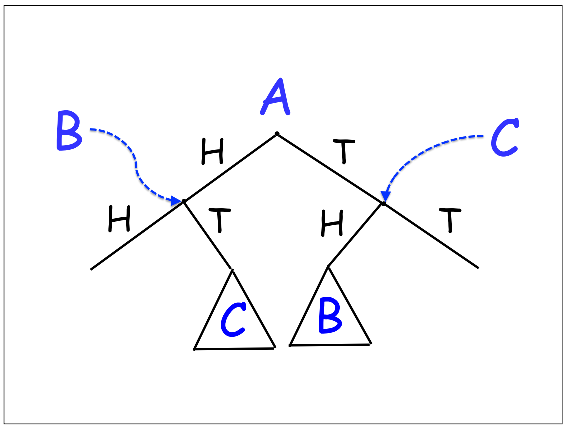


Figure 19.10 Outcome Tree for Flipping Until HH or TT

So we’re interested in finding $e(D)$.

Write a small system of equations involving $e(D)$, $e(B)$, and $e(C)$ that could be solved to find $e(D)$. *You do **not** need to solve the equations.*

Problem 19.17.
A coin with probability p of flipping Heads and probability $q ::= 1 - p$ of flipping tails is repeatedly flipped until two consecutive flips match—that is, until HH or TT occurs. The outcome tree A for this setup is illustrated in Figure 19.10.

Let $e(T)$ be the expected number of flips starting at the root of subtree T of A. So we’re interested in finding $e(A)$.

Write a small system of equations involving $e(A)$, $e(B)$, and $e(C)$ that could be solved to find $e(A)$. *You do **not** need to solve the equations.*

Homework Problems

Problem 19.18.
We are given a random vector of n distinct numbers. We then determine the maximum of these numbers using the following procedure:

Pick the first number. Call it the *current maximum.* Go through the rest of the vector (in order) and each time we come across a number (call it x) that exceeds our current maximum, we update the current maximum with x.

What is the expected number of times we update the current maximum?

Hint: Let X_i be the indicator variable for the event that the ith element in the vector is larger than all the previous elements.

Problem 19.19.
A fair six-sided die is repeatedly thrown until a six appears. We are interested in the expected time for a six to appear under certain conditions.

A natural probability space $\mathcal{S}$ modelling this situation is the set of finite strings $s \in [1..5]^*6$ of integers from one to six that end at the first occurrence of a six, with $\Pr[s] ::= (1/6)^{|s|}$. Let T be the random variable equal to the number of throws until six appears, namely, $T(s)$ is the length $|s|$ of s.

(a) What is the expected time $\text{Ex}[T]$ till a six is thrown?

Let V be the event that all the dice throws are eVen. That is, $V = \{2, 4\}^*6$ is the event that all throws are 2’s and 4’s until the first 6.

(b) Prove that $\Pr[V] = 1/4$.

Hint: $V = 2V \cup 4V \cup \{6\}$.

(c) Use the definition of $\text{Ex}[T \mid V]$ as a sum over $s \in V$ to compute the expected time $\text{Ex}[T \mid V]$ till a six is thrown given that all throws are even.

(d) Given that all throws are even, the only possible throws are two, four and six, so we might as well just consider a three-sided die with sides two, four and six. By Mean Time to Failure, the expected time till a six is thrown by a three-sided die is $1/(1/3) = 3$, so $\text{Ex}[T \mid V] = 3$, contradicting part (c)! Explain.[8]

Problems for Section 19.6

Class Problems

Problem 19.20.
You have a biased coin with nonzero probability $p < 1$ of tossing a Head. You toss until a Head comes up. Then, similar to the example in Section 19.6, you keep tossing until you get another Head preceded by a run of consecutive Tails

[8] If you’re thrown by this, you are not alone. There are several websites devoted to explanations of this seductive problem. In fact, when it came up at the MIT Theory of Computation faculty lunch in April 2018, several attendees confidently defended the mistaken reasoning.

whose length is within 10 of your original run. That is, if you began by tossing k tails followed by a Head, then you continue tossing until you get a run of at least $\max\{k - 10, 0\}$ consecutive Tails.

(a) Let H be the number of Heads that you toss until you get the required run of Tails. Prove that the expected value of H is infinite.

(b) Let $r < 1$ be a positive real number. Instead of waiting for a run of Tails of length $k - 10$ when your original run was length k, just wait for a run of length at least rk. Show that in this case, the expected number of Heads is finite.

Exam Problems

Problem 19.21.
You have a random process for generating a positive integer K. The behavior of your process each time you use it is (mutually) independent of all its other uses. You use your process to generate an integer, and then use your procedure repeatedly until you generate an integer as big as your first one. Let R be the number of additional integers you have to generate.

(a) State and briefly explain a simple closed formula for $\text{Ex}[R \mid K = k]$ in terms of $\Pr[K \geq k]$.

Suppose $\Pr[K = k] = \Theta(k^{-4})$.

(b) Show that $\Pr[K \geq k] = \Theta(k^{-3})$.

(c) Show that $\text{Ex}[R]$ is infinite.

Problems for Section 19.6

Practice Problems

Problem 19.22.
MIT students sometimes delay doing laundry until they finish their problem sets. Assume all random values described below are mutually independent.

(a) A *busy* student must complete 3 problem sets before doing laundry. Each problem set requires 1 day with probability $2/3$ and 2 days with probability $1/3$. Let B be the number of days a busy student delays laundry. What is $\text{Ex}[B]$?

Example: If the first problem set requires 1 day and the second and third problem sets each require 2 days, then the student delays for $B = 5$ days.

(b) A *relaxed* student rolls a fair, 6-sided die in the morning. If he rolls a 1, then he does his laundry immediately (with zero days of delay). Otherwise, he delays for one day and repeats the experiment the following morning. Let R be the number of days a relaxed student delays laundry. What is $\text{Ex}[R]$?

Example: If the student rolls a 2 the first morning, a 5 the second morning, and a 1 the third morning, then he delays for $R = 2$ days.

(c) Before doing laundry, an *unlucky* student must recover from illness for a number of days equal to the product of the numbers rolled on two fair, 6-sided dice. Let U be the expected number of days an unlucky student delays laundry. What is $\text{Ex}[U]$?

Example: If the rolls are 5 and 3, then the student delays for $U = 15$ days.

(d) A student is *busy* with probability $1/2$, *relaxed* with probability $1/3$, and *unlucky* with probability $1/6$. Let D be the number of days the student delays laundry. What is $\text{Ex}[D]$?

Problem 19.23.

Each Math for Computer Science final exam will be graded according to a rigorous procedure:

- With probability $4/7$ the exam is graded by a *TA*,with probability $2/7$ it is graded by a *lecturer*, and with probability $1/7$, it is accidentally dropped behind the radiator and arbitrarily given a score of 84.
- TAs score an exam by scoring each problem individually and then taking the sum.
  - There are ten true/false questions worth 2 points each. For each, full credit is given with probability $3/4$, and no credit is given with probability $1/4$.
  - There are four questions worth 15 points each. For each, the score is determined by rolling two fair dice, summing the results, and adding 3.
  - The single 20 point question is awarded either 12 or 18 points with equal probability.

- Lecturers score an exam by rolling a fair die twice, multiplying the results, and then adding a "general impression"score.
  - With probability $4/10$, the general impression score is 40.
  - With probability $3/10$, the general impression score is 50.
  - With probability $3/10$, the general impression score is 60.

Assume all random choices during the grading process are independent.

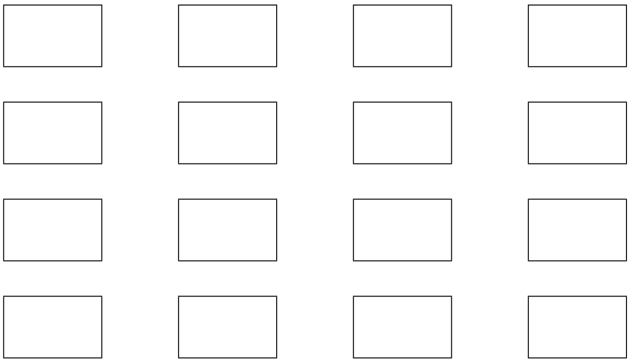
(a) What is the expected score on an exam graded by a TA?

(b) What is the expected score on an exam graded by a lecturer?

(c) What is the expected score on a Math for Computer Science final exam?

Class Problems

Problem 19.24.
A classroom has sixteen desks in a 4×4 arrangement as shown below.



If there is a girl in front, behind, to the left, or to the right of a boy, then the two *flirt*. One student may be in multiple flirting couples; for example, a student in a corner of the classroom can flirt with up to two others, while a student in the center can flirt with as many as four others. Suppose that desks are occupied mutually independently by boys and girls with equal probability. What is the expected number of flirting couples? *Hint:* Linearity.

Problem 19.25.
A *literal* is a propositional variable P or its negation $\overline{P}$, where as usual "$\overline{P}$" abbreviates "NOT(P)." A *3-clause* is an OR of three literals from three different variables. For example,

$$P_1 \text{ OR } P_2 \text{ OR } \overline{P_3}$$

is a 3-clause, but P_1 OR $\overline{P_1}$ OR P_2 is not because P_1 appears twice. A *3-CNF* is a formula that is an AND of 3-clauses. For example,

$$(P_1 \text{ OR } P_2 \text{ OR } \overline{P_3}) \text{ AND } (\overline{P_1} \text{ OR } P_3 \text{ OR } \overline{P_4}) \text{ AND } (P_2 \text{ OR } P_3 \text{ OR } \overline{P_4})$$

is a 3-CNF.

Suppose that G is a 3-CNF with seven 3-clauses. Assign true/false values to the variables in G independently and with equal probability.

(a) What is the probability that the nth clause is true?

(b) What is the expected number of true 3-clauses in G?

(c) Use the fact that the answer to part (b) is greater than six to conclude G must be satisfiable.

Problem 19.26.
A *literal* is a propositional variable or its negation. A *k-clause* is an OR of k literals, with no variable occurring more than once in the clause. For example,

$$P \text{ OR } \overline{Q} \text{ OR } \overline{R} \text{ OR } V,$$

is a 4-clause, but

$$\overline{V} \text{ OR } \overline{Q} \text{ OR } \overline{X} \text{ OR } V,$$

is not, since V appears twice.

Let $\mathcal{S}$ be a set of n distinct k-clauses involving v variables. The variables in different k-clauses may overlap or be completely different, so $k \le v \le nk$.

A random assignment of true/false values will be made independently to each of the v variables, with true and false assignments equally likely. Write formulas in n, k and v in answer to the first two parts below.

(a) What is the probability that any particular k-clause in $\mathcal{S}$ is true under the random assignment?

(b) What is the expected number of true k-clauses in $\mathcal{S}$?

(c) A set of propositions is *satisfiable* iff there is an assignment to the variables that makes all of the propositions true. Use your answer to part (b) to prove that if $n < 2^k$, then $\mathcal{S}$ is satisfiable.

Problem 19.27.
There are n students who are both taking Math for Computer Science (MCS) and Introduction to Signal Processing (SP) this term. To make it easier on themselves, the Professors in charge of these classes have decided to randomly permute their class lists and then assign students grades based on their rank in the permutation (just as many students have suspected). Assume the permutations are equally likely and independent of each other. What is the expected number of students that have in rank in SP that is higher by k than their rank in MCS?

Hint: Let X_r be the indicator variable for the rth ranked student in CS having a rank in SP of at least $r + k$.

Problem 19.28.
A man has a set of n keys, one of which fits the door to his apartment. He tries the keys randomly until he finds the key that fits. Let T be the number of times he tries keys until he finds the right key.

(a) Suppose each time he tries a key that does not fit the door, he simply puts it back. This means he might try the same ill-fitting key several times before he finds the right key. What is $\text{Ex}[T]$?

Hint: Mean time to failure.

Now suppose he throws away each ill-fitting key that he tries. That is, he chooses keys randomly from *among those he has not yet tried.* This way he is sure to find the right key within n tries.

(b) If he hasn't found the right key yet and there are m keys left, what is the probability that he will find the right key on the next try?

(c) Given that he did not find the right key on his first $k - 1$ tries, verify that the probability that he does not find it on the kth trial is given by

$$\Pr\big[T > k \mid T > k - 1\big] = \frac{n - k}{n - (k - 1)}.$$

(d) Prove that

$$\Pr[T > k] = \frac{n - k}{n}. \tag{19.18}$$

Hint: This can be argued directly, but if you don't see how, induction using part (c) will work.

(e) Conclude that in this case

$$\text{Ex}[T] = \frac{n+1}{2}.$$

Problem 19.29.
Justify each line of the following proof that if R and S are *independent* random variables, then

$$\text{Ex}[R \cdot S] = \text{Ex}[R] \cdot \text{Ex}[S].$$

Proof.

$$\begin{aligned}
&\text{Ex}[R \cdot S] \\
&= \sum_{t \in \text{range}(R \cdot S)} t \cdot \Pr[R \cdot S = t] \\
&= \sum_{r \in \text{range}(R),\ s \in \text{range}(S)} rs \cdot \Pr[R = r \text{ and } S = s] \\
&= \sum_{r \in \text{range}(R)} \left(\sum_{s \in \text{range}(S)} rs \cdot \Pr[R = r \text{ and } S = s] \right) \\
&= \sum_{r \in \text{range}(R)} \left(\sum_{s \in \text{range}(S)} rs \cdot \Pr[R = r] \cdot \Pr[S = s] \right) \\
&= \sum_{r \in \text{range}(R)} \left(r \Pr[R = r] \cdot \sum_{s \in \text{range}(S)} s \Pr[S = s] \right) \\
&= \sum_{r \in \text{range}(R)} r \Pr[R = r] \cdot \text{Ex}[S] \\
&= \text{Ex}[S] \cdot \sum_{r \in \text{range}(R)} r \Pr[R = r] \\
&= \text{Ex}[S] \cdot \text{Ex}[R].
\end{aligned}$$

■

Problem 19.30.
A gambler bets on the toss of a fair coin: if the toss is Heads, the gambler gets back the amount he bet along with an additional the amount equal to his bet. Otherwise he loses the amount bet. For example, the gambler bets \$10 and wins, he gets back \$20 for a net profit of \$10. If he loses, he gets back nothing for a net profit of −\$10—that is, a net loss of \$10.

Gamblers often try to develop betting strategies to beat the odds is such a game. A well known strategy of this kind is *bet doubling*, namely, bet \$10 on red, and keep doubling the bet until a red comes up. So if the gambler wins his first \$10 bet, he stops playing and leaves with his \$10 profit. If he loses the first bet, he bets \$20 on the second toss. Now if the second toss is Heads, he gets his \$20 bet plus \$20 back and again walks away with a net profit of $20 - 10 = \$10$. If he loses the second toss, he bets \$40 on the third toss, and so on.

You would think that any such strategy will be doomed: in a fair game your expected win by definition is zero, so no strategy should have nonzero expectation. We can make this reasoning more precise as follows:

> Let W_n be a random variable equal to the amount won in the nth coin toss. So with the bet doubling strategy starting with a \$10 bet, $W_1 = \pm 10$ with equal probability. If the betting ends before the nth bet, define $W_n = 0$. So W_2 is zero with probability 1/2, is 10 with probability 1/4, and is -10 with probability 1/4. Now letting W be the amount won when the gambler stops betting, we have
>
> $$W = W_1 + W_2 + \cdots + W_n + \cdots .$$
>
> Furthermore, since each toss is fair,
>
> $$\text{Ex}[W_n] = 0$$
>
> for all $n > 0$. Now by linearity of expectation, we have
>
> $$\text{Ex}[W] = \text{Ex}[W_1] + \text{Ex}[W_2] + \cdots + \text{Ex}[W_n] + \cdots = 0 + 0 + \cdots + 0 + \cdots = 0, \tag{19.19}$$
>
> confirming that with fair tosses, the expected win is zero.

But wait a minute!

(a) Explain why the gambler is certain to win eventually if he keeps betting.

(b) Prove that when the gambler finally wins a bet, his net profit is \$10.

(c) Since the gambler’s profit is always \$10 when he wins, and he is certain to win, his expected profit is also \$10. That is

$$\text{Ex}[W] = 10,$$

contradicting (19.19). So what’s wrong with the reasoning that led to the false conclusion (19.19)?

Homework Problems

Problem 19.31.

Applying linearity of expectation to the binomial distribution $f_{n,p}$ immediately yielded the identity 19.13:

$$\text{Ex}[f_{n,p}] ::= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = pn. \tag{19.20}$$

Though it might seem daunting to prove this equation without appeal to linearity, it is, after all, pretty similar to the binomial identity, and this connection leads to an immediate alternative algebraic derivation.

(a) Starting with the binomial identity for $(x+y)^n$, prove that

$$xn(x+y)^{n-1} = \sum_{k=0}^{n} k \binom{n}{k} x^k y^{n-k}. \tag{19.21}$$

(b) Now conclude equation (19.20).

Problem 19.32.

Short-term Capital Management (STCM) wants you to invest in a fund with the following rules: you invest one million dollars in their Forward Looking Internet Package (FLIP). Each year, the money in your FLIP account will double or halve with equal probability, and each year STCM will pay you a dividend equal to 10% of the money in your account.

(a) What is the expected number of dollars in your account at the end of k years? Write a simple formula in terms of k.

Hint: \$1,000,000 is in the account the end of year zero. Let X_i be 2 or $1/2$ depending on what happens to your money at the end of the ith year. So the amount of money in the account at the end of year one is $X_1 \cdot \$1,000,000$ and the dividend paid is $(1/10)X_1 \cdot \$1,000,000$.

(b) Give a closed form numerical expression for the expected total number of dollars in dividend payments you will receive by the end of the 10th year. You do *not* need to evaluate your expression.

(c) Adam Smith does his own analysis of your account. He lets $Y_i = 1$ if the money doubles at the end of year i and $Y_i = -1$ otherwise. Then the money in your account after year k is

$$10^6 2^{Y_1} 2^{Y_2} \cdots 2^{Y_k} = 10^6 2^{Y_1+Y_2+\cdots+Y_k}.$$

But $\text{Ex}[Y_i] = 0$, so

$$2^{\text{Ex}[Y_1+Y_2+\cdots+Y_k]} = 2^{\text{Ex}[Y_1]+\text{Ex}[Y_2]+\cdots+\text{Ex}[Y_k]} = 2^{k\cdot 0} = 2^0 = 1.$$

In other words, the expected amount of money in your account forever remains the same as your original investment.

What is wrong with Adam Smith's analysis?

Problem 19.33.
A coin will be flipped repeatedly until the sequence TTH (tail/tail/head) comes up. Successive flips are independent, and the coin has probability p of coming up heads. Let $N_{\texttt{TTH}}$ be the number of coin flips until TTH first appears. What value of p minimizes $\text{Ex}[N_{\texttt{TTH}}]$?

Problem 19.34.
(A true story from World War Two.)

The army needs to test n soldiers for a disease. There is a blood test that accurately determines when a blood sample contains blood from a diseased soldier. The army presumes, based on experience, that the fraction of soldiers with the disease is approximately equal to some small number p.

Approach (1) is to test blood from each soldier individually; this requires n tests. Approach (2) is to randomly group the soldiers into g groups of k soldiers, where $n = gk$. For each group, blend the k blood samples of the people in the group, and test the blended sample. If the group-blend is free of the disease, we are done with that group after one test. If the group-blend tests positive for the disease, then someone in the group has the disease, and we to test all the people in the group for a total of $k + 1$ tests on that group.

Since the groups are chosen randomly, each soldier in the group has the disease with probability p, and it is safe to assume that whether one soldier has the disease is independent of whether the others do.

(a) What is the expected number of tests in Approach (2) as a function of the number of soldiers n, the disease fraction p, and the group size k?

(b) Show how to choose k so that the expected number of tests using Approach (2) is approximately $n\sqrt{p}$. *Hint:* Since p is small, you may assume that $(1-p)^k \approx 1$ and $\ln(1-p) \approx -p$.

(c) What fraction of the work does Approach (2) expect to save over Approach (1) in a million-strong army of whom approximately 1% are diseased?

(d) Can you come up with a better scheme by using multiple levels of grouping, that is, groups of groups?

Problem 19.35.

A wheel-of-fortune has the numbers from 1 to $2n$ arranged in a circle. The wheel has a spinner, and a spin randomly determines the two numbers at the opposite ends of the spinner. How would you arrange the numbers on the wheel to maximize the expected value of:

(a) the sum of the numbers chosen? What is this maximum?

(b) the product of the numbers chosen? What is this maximum?

Hint: For part (b), verify that the sum of the products of numbers oppposite each other is maximized when successive integers are on the opposite ends of the spinner, that is, 1 is opposite 2, 3 is opposite 4, 5 is opposite 6, ...

Problem 19.36.

Let R and S be independent random variables, and f and g be any functions such that $\text{domain}(f) = \text{codomain}(R)$ and $\text{domain}(g) = \text{codomain}(S)$. Prove that $f(R)$ and $g(S)$ are also independent random variables.

Hint: The event $[f(R) = a]$ is the disjoint union of all the events $[R = r]$ for r such that $f(r) = a$.

Problem 19.37.

Peeta bakes between 1 and $2n$ loaves of bread to sell every day. Each day he rolls a fair, n-sided die to get a number from 1 to n, then flips a fair coin. If the coin is heads, he bakes m loaves of bread , where m is the number on the die that day, and if the coin is tails, he bakes $2m$ loaves.

(a) For any positive integer $k \leq 2n$, what is the probability that Peeta will make k loaves of bread on any given day?

Hint: Express your solution by cases.

(b) What is the expected number of loaves that Peeta would bake on any given day?

(c) Continuing this process, Peeta bakes bread every day for 30 days. What is the expected total number of loaves that Peeta would bake?

Exam Problems

Problem 19.38.
A box initially contains n balls, all colored black. A ball is drawn from the box at random.

- If the drawn ball is black, then a biased coin with probability, $p > 0$, of coming up heads is flipped. If the coin comes up heads, a white ball is put into the box; otherwise the black ball is returned to the box.
- If the drawn ball is white, then it is returned to the box.

This process is repeated until the box contains n white balls.

Let D be the number of balls drawn until the process ends with the box full of white balls. Prove that $\text{Ex}[D] = nH_n/p$, where H_n is the nth Harmonic number.

Hint: Let D_i be the number of draws after the ith white ball until the draw when the $(i+1)$st white ball is put into the box.

Problem 19.39.
A gambler bets \$10 on “red” at a roulette table (the odds of red are 18/38, slightly less than even) to win \$10. If he wins, he gets back twice the amount of his bet, and he quits. Otherwise, he doubles his previous bet and continues.

For example, if he loses his first two bets but wins his third bet, the total spent on his three bets is $10 + 20 + 40$ dollars, but he gets back $2 \cdot 40$ dollars after his win on the third bet, for a net profit of \$10.

(a) What is the expected number of bets the gambler makes before he wins?

(b) What is his probability of winning?

(c) What is his expected final profit (amount won minus amount lost)?

(d) You can beat a biased game by bet doubling, but bet doubling is not feasible because it requires an infinite bankroll. Verify this by proving that the expected size of the gambler’s last bet is infinite.

Problem 19.40.
Six pairs of cards with ranks 1–6 are shuffled and laid out in a row, for example,

$$\boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{3}\,\boxed{4}\,\boxed{6}\,\boxed{1}\,\boxed{4}\,\boxed{5}\,\boxed{5}\,\boxed{2}\,\boxed{6}$$

In this case, there are two adjacent pairs with the same value, the two 3’s and the two 5’s. What is the expected number of adjacent pairs with the same value?

Problem 19.41.
There are six kinds of cards, three of each kind, for a total of eighteen cards. The cards are randonly shuffled and laid out in a row, for example,

$$\boxed{1}\,\boxed{2}\,\boxed{5}\,\boxed{5}\,\boxed{5}\,\boxed{1}\,\boxed{4}\,\boxed{6}\,\boxed{2}\,\boxed{6}\boxed{6}\,\boxed{2}\,\boxed{1}\,\boxed{4}\,\boxed{3}\,\boxed{3}\,\boxed{3}\,\boxed{4}$$

In this case, there are two adjacent triples of the same kind, the three 3’s and the three 5’s.

(a) Derive a formula for the probability that the 4th, 5th, and 6th consecutive cards will be the same kind—that is, all 1’s or all 2’s or. . . all 6’s?

(b) Let $p ::= \Pr[\text{4th, 5th and 6th cards match}]$—that is, p is the correct answer to part (a). Write a simple formula for the expected number of matching triples in terms of p.

20 Deviation from the Mean

In the previous chapter, we took it for granted that expectation is useful and developed a bunch of techniques for calculating expected values. But why should we care about this value? After all, a random variable may never take a value anywhere near its expectation.

The most important reason to care about the mean value comes from its connection to estimation by sampling. For example, suppose we want to estimate the average age, income, family size, or other measure of a population. To do this, we determine a random process for selecting people—say, throwing darts at census lists. This process makes the selected person's age, income, and so on into a random variable whose *mean* equals the *actual average* age or income of the population. So, we can select a random sample of people and calculate the average of people in the sample to estimate the true average in the whole population. But when we make an estimate by repeated sampling, we need to know how much confidence we should have that our estimate is OK, and how large a sample is needed to reach a given confidence level. The issue is fundamental to all experimental science. Because of random errors—*noise*—repeated measurements of the same quantity rarely come out exactly the same. Determining how much confidence to put in experimental measurements is a fundamental and universal scientific issue. Technically, judging sampling or measurement accuracy reduces to finding the probability that an estimate *deviates* by a given amount from its expected value.

Another aspect of this issue comes up in engineering. When designing a sea wall, you need to know how strong to make it to withstand tsunamis for, say, at least a century. If you're assembling a computer network, you might need to know how many component failures it should tolerate to likely operate without maintenance for at least a month. If your business is insurance, you need to know how large a financial reserve to maintain to be nearly certain of paying benefits for, say, the next three decades. Technically, such questions come down to finding the probability of *extreme* deviations from the mean.

This issue of *deviation from the mean* is the focus of this chapter.

20.1 Markov's Theorem

Markov's theorem gives a generally coarse estimate of the probability that a random variable takes a value *much larger* than its mean. It is an almost trivial result by itself, but it actually leads fairly directly to much stronger results.

The idea behind Markov's Theorem can be explained by considering the quantity known as *intelligence quotient*, IQ, which remains in wide use despite doubts about its legitimacy. IQ was devised so that its average measurement would be 100. This immediately implies that at most 1/3 of the population can have an IQ of 300 or more, because if more than a third had an IQ of 300, then the average would have to be *more* than $(1/3) \cdot 300 = 100$. So, the probability that a randomly chosen person has an IQ of 300 or more is at most 1/3. By the same logic, we can also conclude that at most 2/3 of the population can have an IQ of 150 or more.

Of course, these are not very strong conclusions. No IQ of over 300 has ever been recorded; and while many IQ's of over 150 have been recorded, the fraction of the population that actually has an IQ that high is very much smaller than 2/3. But though these conclusions are weak, we reached them using just the fact that the average IQ is 100—along with another fact we took for granted, that IQ is never negative. Using only these facts, we can't derive smaller fractions, because there are nonnegative random variables with mean 100 that achieve these fractions. For example, if we choose a random variable equal to 300 with probability 1/3 and 0 with probability 2/3, then its mean is 100, and the probability of a value of 300 or more really is 1/3.

Theorem 20.1.1 (Markov's Theorem)**.** *If R is a nonnegative random variable, then for all $x > 0$*

$$\Pr[R \geq x] \leq \frac{\operatorname{Ex}[R]}{x}. \tag{20.1}$$

Proof. Let I_x be the indicator variable for the event $[R \geq x]$. Then

$$x I_x \leq R$$

holds for all values of R since $R \geq 0$. Taking expectations of both sides yields

$$x \Pr[R \geq x] \leq \operatorname{Ex}[R],$$

and then dividing both sides of this inequality by x gives (20.1). ■

Our focus is deviation from the mean, so it's useful to rephrase Markov's Theorem this way:

Corollary 20.1.2. *If R is a nonnegative random variable, then for all $c \geq 1$*

$$\Pr[R \geq c \cdot \operatorname{Ex}[R]\,] \leq \frac{1}{c}. \tag{20.2}$$

This Corollary follows immediately from Markov's Theorem(20.1.1) by letting x be $c \cdot \operatorname{Ex}[R]$.

20.1.1 Applying Markov's Theorem

Let's go back to the Hat-Check problem of Section 19.5.2. Now we ask what the probability is that x or more men get the right hat, this is, what the value of $\Pr[G \geq x]$ is.

We can compute an upper bound with Markov's Theorem. Since we know $\text{Ex}[G] = 1$, Markov's Theorem implies

$$\Pr[G \geq x] \leq \frac{\text{Ex}[G]}{x} = \frac{1}{x}.$$

For example, there is no better than a 20% chance that 5 men get the right hat, regardless of the number of people at the dinner party.

The Chinese Appetizer problem is similar to the Hat-Check problem. In this case, n people are eating different appetizers arranged on a circular, rotating Chinese banquet tray. Someone then spins the tray so that each person receives a random appetizer. What is the probability that everyone gets the same appetizer as before?

There are n equally likely orientations for the tray after it stops spinning. Everyone gets the right appetizer in just one of these n orientations. Therefore, the correct answer is $1/n$.

But what probability do we get from Markov's Theorem? Let the random variable R be the number of people that get the right appetizer. Then of course $\text{Ex}[R] = 1$, so applying Markov's Theorem, we find:

$$\Pr[R \geq n] \leq \frac{\text{Ex}[R]}{n} = \frac{1}{n}.$$

So for the Chinese appetizer problem, Markov's Theorem is precisely right!

Unfortunately, Markov's Theorem is not always so accurate. For example, it gives the same $1/n$ upper limit for the probability that everyone gets their own hat back in the Hat-Check problem, where the probability is actually $1/(n!)$. So for Hat-Check, Markov's Theorem gives a probability bound that is way too large.

20.1.2 Markov's Theorem for Bounded Variables

Suppose we learn that the average IQ among MIT students is 150 (which is not true, by the way). What can we say about the probability that an MIT student has an IQ of more than 200? Markov's theorem immediately tells us that no more than $150/200$ or $3/4$ of the students can have such a high IQ. Here, we simply applied Markov's Theorem to the random variable R equal to the IQ of a random MIT student to conclude:

$$\Pr[R > 200] \leq \frac{\text{Ex}[R]}{200} = \frac{150}{200} = \frac{3}{4}.$$

But let's observe an additional fact (which may be true): no MIT student has an IQ less than 100. This means that if we let $T ::= R - 100$, then T is nonnegative and $\text{Ex}[T] = 50$, so we can apply Markov's Theorem to T and conclude:

$$\Pr[R > 200] = \Pr[T > 100] \leq \frac{\text{Ex}[T]}{100} = \frac{50}{100} = \frac{1}{2}.$$

So only half, not 3/4, of the students can be as amazing as they think they are. A bit of a relief!

In fact, we can get better bounds applying Markov's Theorem to $R - b$ instead of R for any lower bound b on R (see Problem 20.3). Similarly, if we have any upper bound u on a random variable S, then $u - S$ will be a nonnegative random variable, and applying Markov's Theorem to $u - S$ will allow us to bound the probability that S is much *less* than its expectation.

20.2 Chebyshev's Theorem

We've seen that Markov's Theorem can give a better bound when applied to $R - b$ rather than R. More generally, a good trick for getting stronger bounds on a random variable R out of Markov's Theorem is to apply the theorem to some cleverly chosen function of R. Choosing functions that are powers of the absolute value of R turns out to be especially useful. In particular, since $|R|^z$ is nonnegative for any real number z, Markov's inequality also applies to the event $[\,|R|^z \geq x^z]$. But for positive $x, z > 0$ this event is equivalent to the event $[\,|R| \geq x]$ for , so we have:

Lemma 20.2.1. *For any random variable R and positive real numbers x, z,*

$$\Pr[|R| \geq x] \leq \frac{\text{Ex}[\,|R|^z]}{x^z}.$$

Rephrasing (20.2.1) in terms of $|R - \text{Ex}[R]\,|$, the random variable that measures R's deviation from its mean, we get

$$\Pr[\,|R - \text{Ex}[R]\,| \geq x] \leq \frac{\text{Ex}[(|R - \text{Ex}[R]|)^z]}{x^z}. \tag{20.3}$$

When z is positive and even, $(R - \text{Ex}[R])^z$ is nonnegative, so the absolute value on the right-hand side of the inequality (20.3) is redundant. The case when $z = 2$ turns out to be so important that the numerator of the right-hand side has been given a name:

Definition 20.2.2. The *variance* of a random variable R is:

$$\text{Var}[R] ::= \text{Ex}\left[(R - \text{Ex}[R])^2\right].$$

Variance is also known as *mean square deviation*.

The restatement of (20.3) for $z = 2$ is known as *Chebyshev's Theorem.*[1]

Theorem 20.2.3 (Chebyshev)**.** *Let R be a random variable and $x \in \mathbb{R}^+$. Then*

$$\Pr[|R - \text{Ex}[R]| \geq x] \leq \frac{\text{Var}[R]}{x^2}.$$

The expression $\text{Ex}[(R - \text{Ex}[R])^2]$ for variance is a bit cryptic; the best approach is to work through it from the inside out. The innermost expression $R - \text{Ex}[R]$ is precisely the deviation of R above its mean. Squaring this, we obtain $(R-\text{Ex}[R])^2$. This is a random variable that is near 0 when R is close to the mean and is a large positive number when R deviates far above or below the mean. So if R is always close to the mean, then the variance will be small. If R is often far from the mean, then the variance will be large.

20.2.1 Variance in Two Gambling Games

The relevance of variance is apparent when we compare the following two gambling games.

Game A: We win \$2 with probability $2/3$ and lose \$1 with probability $1/3$.

Game B: We win \$1002 with probability $2/3$ and lose \$2001 with probability $1/3$.

Which game is better financially? We have the same probability, 2/3, of winning each game, but that does not tell the whole story. What about the expected return for each game? Let random variables A and B be the payoffs for the two games. For example, A is 2 with probability 2/3 and -1 with probability 1/3. We can compute the expected payoff for each game as follows:

$$\text{Ex}[A] = 2 \cdot \frac{2}{3} + (-1) \cdot \frac{1}{3} = 1,$$
$$\text{Ex}[B] = 1002 \cdot \frac{2}{3} + (-2001) \cdot \frac{1}{3} = 1.$$

The expected payoff is the same for both games, but the games are very different. This difference is not apparent in their expected value, but is captured by variance.

[1] There are Chebyshev Theorems in several other disciplines, but Theorem 20.2.3 is the only one we'll refer to.

We can compute the Var[A] by working "from the inside out" as follows:

$$\begin{aligned}
A - \text{Ex}[A] &= \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ -2 & \text{with probability } \frac{1}{3} \end{cases} \\
(A - \text{Ex}[A])^2 &= \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ 4 & \text{with probability } \frac{1}{3} \end{cases} \\
\text{Ex}[(A - \text{Ex}[A])^2] &= 1 \cdot \frac{2}{3} + 4 \cdot \frac{1}{3} \\
\text{Var}[A] &= 2.
\end{aligned}$$

Similarly, we have for Var[B]:

$$\begin{aligned}
B - \text{Ex}[B] &= \begin{cases} 1001 & \text{with probability } \frac{2}{3} \\ -2002 & \text{with probability } \frac{1}{3} \end{cases} \\
(B - \text{Ex}[B])^2 &= \begin{cases} 1,002,001 & \text{with probability } \frac{2}{3} \\ 4,008,004 & \text{with probability } \frac{1}{3} \end{cases} \\
\text{Ex}[(B - \text{Ex}[B])^2] &= 1,002,001 \cdot \frac{2}{3} + 4,008,004 \cdot \frac{1}{3} \\
\text{Var}[B] &= 2,004,002.
\end{aligned}$$

The variance of Game A is 2 and the variance of Game B is more than two million! Intuitively, this means that the payoff in Game A is usually close to the expected value of \$1, but the payoff in Game B can deviate very far from this expected value.

High variance is often associated with high risk. For example, in ten rounds of Game A, we expect to make \$10, but could conceivably lose \$10 instead. On the other hand, in ten rounds of game B, we also expect to make \$10, but could actually lose more than \$20,000!

20.2.2 Standard Deviation

In Game B above, the deviation from the mean is 1001 in one outcome and -2002 in the other. But the variance is a whopping 2,004,002. The happens because the "units" of variance are wrong: if the random variable is in dollars, then the expectation is also in dollars, but the variance is in square dollars. For this reason, people often describe random variables using *standard deviation* instead of variance.

Definition 20.2.4. The *standard deviation* σ_R of a random variable R is the square root of the variance:

$$\sigma_R ::= \sqrt{\text{Var}[R]} = \sqrt{\text{Ex}[(R - \text{Ex}[R])^2]}.$$

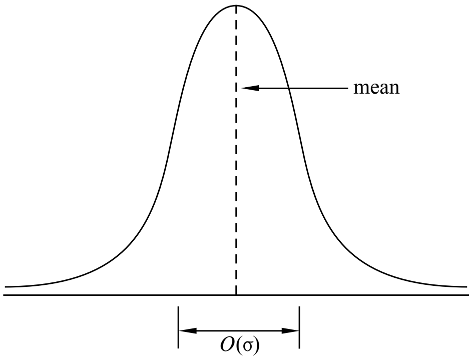


Figure 20.1 The standard deviation of a distribution indicates how wide the “main part” of it is.

So the standard deviation is the square root of the mean square deviation, or the *root mean square* for short. It has the same units—dollars in our example—as the original random variable and as the mean. Intuitively, it measures the average deviation from the mean, since we can think of the square root on the outside as canceling the square on the inside.

Example 20.2.5. The standard deviation of the payoff in Game B is:

$$\sigma_B = \sqrt{\text{Var}[B]} = \sqrt{2,004,002} \approx 1416.$$

The random variable B actually deviates from the mean by either positive 1001 or negative 2002, so the standard deviation of 1416 describes this situation more closely than the value in the millions of the variance.

For bell-shaped distributions like the one illustrated in Figure 20.1, the standard deviation measures the “width” of the interval in which values are most likely to fall. This can be more clearly explained by rephrasing Chebyshev’s Theorem in terms of standard deviation, which we can do by substituting $x = c\sigma_R$ in (20.1):

Corollary 20.2.6. *Let R be a random variable, and let c be a positive real number.*

$$\Pr[|R - \text{Ex}[R]| \geq c\sigma_R] \leq \frac{1}{c^2}. \tag{20.4}$$

Now we see explicitly how the “likely” values of R are clustered in an $O(\sigma_R)$-sized region around $\text{Ex}[R]$, confirming that the standard deviation measures how spread out the distribution of R is around its mean.

The IQ Example

The standard standard deviation of IQ's regularly turns out to be about 15 even across different populations. This additional fact along with the national average IQ being 100 allows a better determination of the occurrence of IQ's of 300 or more.

Let the random variable R be the IQ of a random person. So $\text{Ex}[R] = 100$, $\sigma_R = 15$ and R is nonnegative. We want to compute $\Pr[R \geq 300]$.

We have already seen that Markov's Theorem 20.1.1 gives a coarse bound, namely,

$$\Pr[R \geq 300] \leq \frac{1}{3}.$$

Now we apply Chebyshev's Theorem to the same problem:

$$\Pr[R \geq 300] = \Pr[|R - 100| \geq 200] \leq \frac{\text{Var}[R]}{200^2} = \frac{15^2}{200^2} \approx \frac{1}{178}.$$

So Chebyshev's Theorem implies that at most one person in 178 has an IQ of 300 or more. We have gotten a much tighter bound using additional information—the variance of R—than we could get knowing only the expectation.

20.3 Properties of Variance

Variance is the average *of the square* of the distance from the mean. For this reason, variance is sometimes called the "mean square deviation." Then we take its square root to get the standard deviation—which in turn is called "root mean square deviation."

But why bother squaring? Why not study the actual distance from the mean, namely, the absolute value of $R - \text{Ex}[R]$, instead of its root mean square? The answer is that variance and standard deviation have useful properties that make them much more important in probability theory than average absolute deviation. In this section, we'll describe some of those properties. In the next section, we'll see why these properties are important.

20.3.1 A Formula for Variance

Applying linearity of expectation to the formula for variance yields a convenient alternative formula.

Lemma 20.3.1.

$$\text{Var}[R] = \text{Ex}[R^2] - \text{Ex}^2[R],$$

for any random variable R.

Here we use the notation $\text{Ex}^2[R]$ as shorthand for $(\text{Ex}[R])^2$.

Proof. Let $\mu = \text{Ex}[R]$. Then

$$\begin{aligned}
\text{Var}[R] &= \text{Ex}[(R - \text{Ex}[R])^2] && \text{(Def 20.2.2 of variance)} \\
&= \text{Ex}[(R - \mu)^2] && \text{(def of } \mu) \\
&= \text{Ex}[R^2 - 2\mu R + \mu^2] \\
&= \text{Ex}[R^2] - 2\mu \,\text{Ex}[R] + \mu^2 && \text{(linearity of expectation)} \\
&= \text{Ex}[R^2] - 2\mu^2 + \mu^2 && \text{(def of } \mu) \\
&= \text{Ex}[R^2] - \mu^2 \\
&= \text{Ex}[R^2] - \text{Ex}^2[R]. && \text{(def of } \mu)
\end{aligned}$$

■

A simple and very useful formula for the variance of an indicator variable is an immediate consequence.

Corollary 20.3.2. *If B is a Bernoulli variable where $p ::= \Pr[B = 1]$ and $q ::= 1 - p$, then*

$$\text{Var}[B] = p - p^2 = pq. \tag{20.5}$$

Proof. By Lemma 19.4.2, $\text{Ex}[B] = p$. But B only takes values 0 and 1, so $B^2 = B$ and equation (20.5) follows immediately from Lemma 20.3.1. ■

20.3.2 Variance of Time to Failure

According to Section 19.4.6, the mean time to failure is $1/p$ for a process that fails during any given hour with probability p. What about the variance?

By Lemma 20.3.1,

$$\text{Var}[C] = \text{Ex}[C^2] - (1/p)^2 \tag{20.6}$$

so all we need is a formula for $\text{Ex}[C^2]$.

Now $\text{Ex}[C^2] ::= \sum_{i \geq 1} i^2 q^{i-1} p$ by definition, and we could evaluate this series using methods from Chapter 14 or 16.

A simpler alternative appeals to conditional expectation much as we did in Section 19.4.6 to derive the formula for mean time to failure. Namely, the expected value of C^2 is the probability p of failure in the first hour times 1^2, plus the probability q of non-failure in the first hour times the expected value of $(C + 1)^2$. So

$$\begin{aligned}
\text{Ex}[C^2] &= p \cdot 1^2 + q \,\text{Ex}[(C + 1)^2] \\
&= p + q \left(\text{Ex}[C^2] + \frac{2}{p} + 1 \right) \\
&= p + q \,\text{Ex}[C^2] + q \left(\frac{2}{p} + 1 \right), \quad \text{so} \\
p \,\text{Ex}[C^2] &= p + q \left(\frac{2}{p} + 1 \right) \\
&= \frac{p^2 + q(2 + p)}{p} \quad \text{and} \\
\text{Ex}[C^2] &= \frac{2 - p}{p^2}
\end{aligned}$$

Combining this with (20.6) proves

Lemma 20.3.3. *If failures occur with probability p independently at each step, and C is the number of steps until the first failure,*[2] *then*

$$\text{Var}[C] = \frac{q}{p^2}. \tag{20.7}$$

20.3.3 Dealing with Constants

It helps to know how to calculate the variance of $aR + b$:

Theorem 20.3.4. *[*Square Multiple Rule *for Variance] Let R be a random variable and a a constant. Then*

$$\text{Var}[aR] = a^2 \,\text{Var}[R]. \tag{20.8}$$

Proof. Beginning with the definition of variance and repeatedly applying linearity

[2] That is, C has the geometric distribution with parameter p according to Definition 19.4.7.

of expectation, we have:

$$\begin{aligned}
\text{Var}[aR] &::= \text{Ex}[(aR - \text{Ex}[aR])^2] \\
&= \text{Ex}[(aR)^2 - 2aR\,\text{Ex}[aR] + \text{Ex}^2[aR]] \\
&= \text{Ex}[(aR)^2] - \text{Ex}[2aR\,\text{Ex}[aR]] + \text{Ex}^2[aR] \\
&= a^2\,\text{Ex}[R^2] - 2\,\text{Ex}[aR]\,\text{Ex}[aR] + \text{Ex}^2[aR] \\
&= a^2\,\text{Ex}[R^2] - a^2\,\text{Ex}^2[R] \\
&= a^2\left(\text{Ex}[R^2] - \text{Ex}^2[R]\right) \\
&= a^2\,\text{Var}[R] && \text{(Lemma 20.3.1)}
\end{aligned}$$

■

It's even simpler to prove that adding a constant does not change the variance, as the reader can verify:

Theorem 20.3.5. *Let R be a random variable, and b a constant. Then*

$$\text{Var}[R + b] = \text{Var}[R]. \tag{20.9}$$

Recalling that the standard deviation is the square root of variance, this implies that the standard deviation of $aR + b$ is simply $|a|$ times the standard deviation of R:

Corollary 20.3.6.

$$\sigma_{(aR+b)} = |a|\,\sigma_R.$$

20.3.4 Variance of a Sum

In general, the variance of a sum is not equal to the sum of the variances, but variances do add for *independent* variables. In fact, *mutual* independence is not necessary: *pairwise* independence will do. This is useful to know because there are some important situations, such as Birthday Matching in Section 17.4, that involve variables that are pairwise independent but not mutually independent.

Theorem 20.3.7. *If R and S are independent random variables, then*

$$\text{Var}[R + S] = \text{Var}[R] + \text{Var}[S]. \tag{20.10}$$

Proof. We may assume that $\text{Ex}[R] = 0$, since we could always replace R by $R - \text{Ex}[R]$ in equation (20.10); likewise for S. This substitution preserves the independence of the variables, and by Theorem 20.3.5, does not change the variances.

But for any variable T with expectation zero, we have $\text{Var}[T] = \text{Ex}[T^2]$, so we need only prove

$$\text{Ex}[(R+S)^2] = \text{Ex}[R^2] + \text{Ex}[S^2]. \tag{20.11}$$

But (20.11) follows from linearity of expectation and the fact that

$$\text{Ex}[RS] = \text{Ex}[R]\,\text{Ex}[S] \tag{20.12}$$

since R and S are independent:

$$\begin{aligned}
\text{Ex}[(R+S)^2] &= \text{Ex}[R^2 + 2RS + S^2] \\
&= \text{Ex}[R^2] + 2\,\text{Ex}[RS] + \text{Ex}[S^2] \\
&= \text{Ex}[R^2] + 2\,\text{Ex}[R]\,\text{Ex}[S] + \text{Ex}[S^2] \qquad \text{(by (20.12))} \\
&= \text{Ex}[R^2] + 2 \cdot 0 \cdot 0 + \text{Ex}[S^2] \\
&= \text{Ex}[R^2] + \text{Ex}[S^2].
\end{aligned}$$

■

It’s easy to see that additivity of variance does not generally hold for variables that are not independent. For example, if $R = S$, then equation (20.10) becomes $\text{Var}[R+R] = \text{Var}[R]+\text{Var}[R]$. By the Square Multiple Rule, Theorem 20.3.4, this holds iff $4\,\text{Var}[R] = 2\,\text{Var}[R]$, which implies that $\text{Var}[R] = 0$. So equation (20.10) fails when $R = S$ and R has nonzero variance.

The proof of Theorem 20.3.7 carries over to the sum of any finite number of variables (Problem 20.19), so we have:

Theorem 20.3.8. *[Pairwise Independent Additivity of Variance] If $R_1, R_2, \dots, R_n$ are* pairwise *independent random variables, then*

$$\text{Var}[R_1 + R_2 + \cdots + R_n] = \text{Var}[R_1] + \text{Var}[R_2] + \cdots + \text{Var}[R_n]. \tag{20.13}$$

Now we have a simple way of computing the variance of a variable J that has an (n,p)-binomial distribution. We know that $J = \sum_{k=1}^{n} I_k$ where the I_k are mutually independent indicator variables with $\Pr[I_k = 1] = p$. The variance of each I_k is pq by Corollary 20.3.2, so by linearity of variance, we have

Lemma 20.3.9 (Variance of the Binomial Distribution)**.** *If J has the (n,p)-binomial distribution, then*

$$\text{Var}[J] = n\,\text{Var}[I_k] = npq. \tag{20.14}$$

20.3.5 Matching Birthdays

We saw in Section 17.4 that in a class of 95 students, it is virtually certain that at least one pair of students will have the same birthday. In fact, several pairs of students are likely to have the same birthday. How many matched birthdays should we expect, and how likely are we to see that many matches in a random group of students?

Having matching birthdays for different pairs of students are *not* mutually independent events. If Alice matches Bob and Alice matches Carol, it's certain that Bob and Carol match as well! So the events that various pairs of students have matching birthdays are not even three-way independent.

But knowing that Alice's birthday matches Bob's tells us nothing about who Carol matches. This means that the events that a pair of people have matching birthdays are pairwise independent (see Problem 19.2). So pairwise independent additivity of variance, Theorem 20.3.8, will allow us to calculate the variance of the number of birthday pairs and then apply Chebyshev's bound to estimate the liklihood of seeing some given number of matching pairs.

In particular, suppose there are n students and d days in the year, and let M be the number of pairs of students with matching birthdays. Namely, let $B_1, B_2, \dots, B_n$ be the birthdays of n independently chosen people, and let $E_{i,j}$ be the indicator variable for the event that the ith and jth people chosen have the same birthdays, that is, the event $[B_i = B_j]$. So in our probability model, the B_i's are mutually independent variables, and the $E_{i,j}$'s are pairwise independent. Also, the expectations of $E_{i,j}$ for $i \neq j$ equals the probability that $B_i = B_j$, namely, $1/d$.

Now the number M of matching pairs of birthdays among the n choices is simply the sum of the $E_{i,j}$'s:

$$M = \sum_{1 \le i < j \le n} E_{i,j}. \tag{20.15}$$

Linearity of expectation make it easy to calculate the expected number of pairs of students with matching birthdays.

$$\text{Ex}[M] = \text{Ex}\left[\sum_{1 \le i < j \le n} E_{i,j}\right] = \sum_{1 \le i < j \le n} \text{Ex}[E_{i,j}] = \binom{n}{2} \cdot \frac{1}{d}.$$

Similarly, pairwise independence makes it easy to calculate the variance.

$$\begin{aligned}
\mathrm{Var}[M] &= \mathrm{Var}\left[\sum_{1\le i<j\le n} E_{i,j}\right] \\
&= \sum_{1\le i<j\le n} \mathrm{Var}[E_{i,j}] && \text{(Theorem 20.3.8)} \\
&= \binom{n}{2}\cdot\frac{1}{d}\left(1-\frac{1}{d}\right). && \text{(Corollary 20.3.2)}
\end{aligned}$$

In particular, for a class of $n = 95$ students with $d = 365$ possible birthdays, we have $\mathrm{Ex}[M] \approx 12.23$ and $\mathrm{Var}[M] \approx 12.23(1-1/365) < 12.2$. So by Chebyshev’s Theorem

$$\Pr[|M - \mathrm{Ex}[M]| \ge x] < \frac{12.2}{x^2}.$$

Letting $x = 7$, we conclude that there is a better than 75% chance that in a class of 95 students, the number of pairs of students with the same birthday will be within 7 of 12.23, that is, between 6 and 19.

20.4 Estimation by Random Sampling

Massachusetts Democrats were astonished in 2010 when their early polls of sample voters showed Republican Scott Brown was favored by a majority of voters and so would win the special election to fill the Senate seat that the late Democrat Teddy Kennedy had occupied for over 40 years. Based on their poll results, they mounted an intense, but ultimately unsuccessful, effort to save the seat for their party.

20.4.1 A Voter Poll

Suppose at some time before the election that p was the fraction of voters favoring Scott Brown. We want to estimate this unknown fraction p. Suppose we have some random process for selecting voters from registration lists that selects each voter with equal probability. We can define an indicator variable K by the rule that $K = 1$ if the random voter most prefers Brown, and $K = 0$ otherwise.

Now to estimate p, we take a large number n of random choices of voters[3] and

[3] We’re choosing a random voter n times *with replacement*. We don’t remove a chosen voter from the set of voters eligible to be chosen later; so we might choose the same voter more than once! We would get a slightly better estimate if we required n *different* people to be chosen, but doing so complicates both the selection process and its analysis for little gain.

count the fraction who favor Brown. That is, we define variables $K_1, K_2, \dots,$ where K_i is interpreted to be the indicator variable for the event that the ith chosen voter prefers Brown. Since our choices are made independently, the K_i's are independent. So formally, we model our estimation process by assuming we have mutually independent indicator variables $K_1, K_2, \dots,$ each with the same probability p of being equal to 1. Now let S_n be their sum, that is,

$$S_n ::= \sum_{i=1}^{n} K_i. \tag{20.16}$$

The variable S_n/n describes the fraction of sampled voters who favor Scott Brown. Most people intuitively, and correctly, expect this sample fraction to give a useful approximation to the unknown fraction p.

So we will use the sample value S_n/n as our *statistical estimate* of p. We know that S_n has a binomial distribution with parameters n and p; we can choose n, but p is unknown.

How Large a Sample?

Suppose we want our estimate to be within 0.04 of the fraction p at least 95% of the time. This means we want

$$\Pr\left[\left|\frac{S_n}{n} - p\right| \le 0.04\right] \ge 0.95\,. \tag{20.17}$$

So we'd better determine the number n of times we must poll voters so that inequality (20.17) will hold. Chebyshev's Theorem offers a simple way to determine such a n.

S_n is binomially distributed. Equation (20.14), combined with the fact that pq is maximized when $p = q$, that is, when $p = 1/2$ (check for yourself!), gives

$$\text{Var}[S_n] = n(pq) \le n \cdot \frac{1}{4} = \frac{n}{4}. \tag{20.18}$$

Next, we bound the variance of S_n/n:

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \text{Var}[S_n] && \text{(Square Multiple Rule for Variance (20.8))}\\
&\le \left(\frac{1}{n}\right)^2 \frac{n}{4} && \text{(by (20.18))}\\
&= \frac{1}{4n} && (20.19)
\end{aligned}$$

Using Chebyshev's bound and (20.19) we have:

$$\Pr\left[\left|\frac{S_n}{n} - p\right| \geq 0.04\right] \leq \frac{\operatorname{Var}[S_n/n]}{(0.04)^2} \leq \frac{1}{4n(0.04)^2} = \frac{156.25}{n} \tag{20.20}$$

To make our our estimate with 95% confidence, we want the right-hand side of (20.20) to be at most 1/20. So we choose n so that

$$\frac{156.25}{n} \leq \frac{1}{20},$$

that is,

$$n \geq 3,125.$$

Section 20.5.2 describes how to get tighter estimates of the tails of binomial distributions that lead to a bound on n that is about four times smaller than the one above. But working through this example using only the variance illustrates an approach to estimation that is applicable to arbitrary random variables, not just binomial variables.

20.4.2 Pairwise Independent Sampling

The reasoning we used above to analyze voter polling and matching birthdays is very similar. We summarize it in slightly more general form with a basic result called the Pairwise Independent Sampling Theorem. In particular, we do not need to restrict ourselves to sums of zero-one valued variables, or to variables with the same distribution. For simplicity, we state the Theorem for pairwise independent variables with possibly different distributions but with the same mean and variance.

Theorem 20.4.1 (Pairwise Independent Sampling)**.** *Let $G_1, \dots, G_n$ be pairwise independent variables with the same mean μ and deviation σ. Define*

$$S_n ::= \sum_{i=1}^{n} G_i. \tag{20.21}$$

Then

$$\Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq x\right] \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.$$

Proof. We observe first that the expectation of S_n/n is μ:

$$\begin{aligned}
\text{Ex}\left[\frac{S_n}{n}\right] &= \text{Ex}\left[\frac{\sum_{i=1}^n G_i}{n}\right] && \text{(def of } S_n\text{)}\\
&= \frac{\sum_{i=1}^n \text{Ex}[G_i]}{n} && \text{(linearity of expectation)}\\
&= \frac{\sum_{i=1}^n \mu}{n}\\
&= \frac{n\mu}{n} = \mu.
\end{aligned}$$

The second important property of S_n/n is that its variance is the variance of G_i divided by n:

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \text{Var}[S_n] && \text{(Square Multiple Rule for Variance (20.8))}\\
&= \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^n G_i\right] && \text{(def of } S_n\text{)}\\
&= \frac{1}{n^2}\sum_{i=1}^n \text{Var}[G_i] && \text{(pairwise independent additivity)}\\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. && (20.22)
\end{aligned}$$

This is enough to apply Chebyshev’s Theorem and conclude:

$$\begin{aligned}
\Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq x\right] &\leq \frac{\text{Var}\left[S_n/n\right]}{x^2}. && \text{(Chebyshev's bound)}\\
&= \frac{\sigma^2/n}{x^2} && \text{(by (20.22))}\\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

■

The Pairwise Independent Sampling Theorem provides a quantitative general statement about how the average of independent samples of a random variable approaches the mean. In particular, it proves what is known as the Law of Large Numbers:[4] by choosing a large enough sample size, we can get arbitrarily accurate estimates of the mean with confidence arbitrarily close to 100%.

[4] This is the *Weak* Law of Large Numbers. As you might suppose, there is also a Strong Law, but it’s outside the scope of 6.042.

Corollary 20.4.2. *[Weak Law of Large Numbers] Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean μ, and the same finite deviation, and let*

$$S_n ::= \frac{\sum_{i=1}^{n} G_i}{n}.$$

Then for every $\epsilon > 0$,

$$\lim_{n \to \infty} \Pr[|S_n - \mu| \leq \epsilon] = 1.$$

20.4.3 Confidence in an Estimation

So Chebyshev’s Bound implies that sampling 3,125 voters will yield a fraction that, 95% of the time, is within 0.04 of the actual fraction of the voting population who prefer Brown.

Notice that the actual size of the voting population was never considered because *it did not matter*. People who have not studied probability theory often insist that the population size should influence the sample size. But our analysis shows that polling a little over 3000 people people is always sufficient, regardless of whether there are ten thousand, or a million, or a billion voters. You should think about an intuitive explanation that might persuade someone who thinks population size matters.

Now suppose a pollster actually takes a sample of 3,125 random voters to estimate the fraction of voters who prefer Brown, and the pollster finds that 1250 of them prefer Brown. It’s tempting, **but sloppy**, to say that this means:

False Claim. *With probability 0.95, the fraction p of voters who prefer Brown is* $1250/3125 \pm 0.04$. *Since* $1250/3125 - 0.04 > 1/3$*, there is a 95% chance that more than a third of the voters prefer Brown to all other candidates.*

As already discussed in Section 18.9, what’s objectionable about this statement is that it talks about the probability or “chance” that a real world fact is true, namely that the actual fraction p of voters favoring Brown is more than 1/3. But p is what it is, and it simply makes no sense to talk about the probability that it is something else. For example, suppose p is actually 0.3; then it’s nonsense to ask about the probability that it is within 0.04 of 1250/3125. It simply isn’t.

This example of voter preference is typical: we want to estimate a fixed, unknown real-world quantity. But *being unknown does not make this quantity a random variable*, so it makes no sense to talk about the probability that it has some property.

A more careful summary of what we have accomplished goes this way:

> We have described a probabilistic procedure for estimating the value of the actual fraction p. The probability that *our estimation procedure* will yield a value within 0.04 of p is 0.95.

This is a bit of a mouthful, so special phrasing closer to the sloppy language is commonly used. The pollster would describe his conclusion by saying that

> At the 95% *confidence level*, the fraction of voters who prefer Brown is $1250/3125 \pm 0.04$.

So confidence levels refer to the results of estimation procedures for real-world quantities. The phrase "confidence level" should be heard as a reminder that some statistical procedure was used to obtain an estimate. To judge the credibility of the estimate, it may be important to examine how well this procedure was performed. More important, the confidence assertion above can be rephrased as

> **Either** the fraction of voters who prefer Brown is $1250/3125 \pm 0.04$
> **or** something unlikely (probability 1/20) happened.

If our experience led us to judge that having the preference fraction actually be in this particular interval was unlikely, then this level of confidence would justifiably remain unconvincing.

20.5 Sums of Random Variables

If all you know about a random variable is its mean and variance, then Chebyshev's Theorem is the best you can do when it comes to bounding the probability that the random variable deviates from its mean. In some cases, however, we know more—for example, that the random variable has a binomial distribution—and then it is possible to prove much stronger bounds. Instead of polynomially small bounds such as $1/c^2$, we can sometimes even obtain exponentially small bounds such as $1/e^c$. As we will soon discover, this is the case whenever the random variable T is the sum of n mutually independent random variables $T_1, T_2, \dots, T_n$ where $0 \le T_i \le 1$. A random variable with a binomial distribution is just one of many examples of such a T.

20.5.1 A Motivating Example

Fussbook is a new social networking site oriented toward unpleasant people. Like all major web services, Fussbook has a load balancing problem: it receives lots of forum posts that computer servers have to process. If any server is assigned more work than it can complete in a given interval, then it is overloaded and system performance suffers. That would be bad, because Fussbook users are *not* a tolerant bunch. So balancing the work load across mutliple servers is vital.

An early idea was to assign each server an alphabetic range of forum topics. (“That oughta work!”, one programmer said.) But after the computer handling the “*pr*ivacy” and “*pr*eferred text editor” threads melted from overload, the drawback of an *ad hoc* approach was clear: it’s easy to miss something that will mess up your plan.

If the length of every task were known in advance, then finding a balanced distribution would be a kind of “bin packing” problem. Such problems are hard to solve exactly, but approximation algorithms can come close. Unfortunately, in this case task lengths are not known in advance, which is typical of workload problems in the real world.

So the load balancing problem seems sort of hopeless, because there is no data available to guide decisions. So the programmers of Fussbook gave up and just randomly assigned posts to computers. Imagine their surprise when the system stayed up and hasn’t crashed yet!

As it turns out, random assignment not only balances load reasonably well, but also permits provable performance guarantees. In general, a randomized approach to a problem is worth considering when a deterministic solution is hard to compute or requires unavailable information.

Specifically, Fussbook receives 24,000 forum posts in every 10-minute interval. Each post is assigned to one of several servers for processing, and each server works sequentially through its assigned tasks. It takes a server an average of $1/4$ second to process a post. Some posts, such as pointless grammar critiques and snide witticisms, are easier, but no post—not even the most protracted harangues—takes more than one full second.

Measuring workload in seconds, this means a server is overloaded when it is assigned more than 600 units of work in a given 600 second interval. Fussbook’s average processing load of $24{,}000 \cdot 1/4 = 6000$ seconds per interval would keep 10 computers running at 100% capacity with perfect load balancing. Surely, more than 10 servers are needed to cope with random fluctuations in task length and imperfect load balance. But would 11 be enough? . . . or 15, 20, 100? We’ll answer that question with a new mathematical tool.

20.5.2 The Chernoff Bound

The Chernoff[5] bound is a hammer that you can use to nail a great many problems. Roughly, the Chernoff bound says that certain random variables are very unlikely to significantly exceed their expectation. For example, if the expected load on a processor is just a bit below its capacity, then that processor is unlikely to be

[5] Yes, this is the same Chernoff who figured out how to beat the state lottery—this guy knows a thing or two.

overloaded, provided the conditions of the Chernoff bound are satisfied.

More precisely, the Chernoff Bound says that *the sum of lots of little, independent, random variables is unlikely to significantly exceed the mean of the sum.* The Markov and Chebyshev bounds lead to the same kind of conclusion but typically provide much weaker bounds. In particular, the Markov and Chebyshev bounds are polynomial, while the Chernoff bound is exponential.

Here is the theorem. The proof will come later in Section 20.5.6.

Theorem 20.5.1 (Chernoff Bound)**.** *Let $T_1, \dots T_n$ be mutually independent random variables such that $0 \le T_i \le 1$ for all i. Let $T = T_1 + \cdots + T_n$. Then for all $c \ge 1$,*

$$\Pr[T \ge c \operatorname{Ex}[T]] \le e^{-\beta(c) \operatorname{Ex}[T]} \tag{20.23}$$

where $\beta(c) ::= c \ln c - c + 1$.

The Chernoff bound applies only to distributions of sums of independent random variables that take on values in the real interval $[0, 1]$. The binomial distribution is the most well-known distribution that fits these criteria, but many others are possible, because the Chernoff bound allows the variables in the sum to have differing, arbitrary, or even unknown distributions over the range $[0, 1]$. Furthermore, there is no direct dependence on either the number of random variables in the sum or their expectations. In short, the Chernoff bound gives strong results for lots of problems based on little information—no wonder it is widely used!

20.5.3 Chernoff Bound for Binomial Tails

The Chernoff bound can be applied in easy steps, though the details can be daunting at first. Let's walk through a simple example to get the hang of it: bounding the probability that the number of heads that come up in 1000 independent tosses of a coin exceeds the expectation by 20% or more. Let T_i be an indicator variable for the event that the ith coin is heads. Then the total number of heads is

$$T = T_1 + \cdots + T_{1000}.$$

The Chernoff bound requires that the random variables T_i be mutually independent and take on values in the range $[0, 1]$. Both conditions hold here. In this example the T_i's only take the two values 0 and 1, since they're indicators.

The goal is to bound the probability that the number of heads exceeds its expectation by 20% or more; that is, to bound $\Pr[T \ge c \operatorname{Ex}[T]]$ where c = 1.2. To that end, we compute $\beta(c)$ as defined in the theorem:

$$\beta(c) = c \ln(c) - c + 1 = 0.0187 \dots.$$

If we assume the coin is fair, then $\text{Ex}[T] = 500$. Plugging these values into the Chernoff bound gives:

$$\begin{aligned} \Pr\left[T \geq 1.2\,\text{Ex}[T]\right] &\leq e^{-\beta(c)\cdot\,\text{Ex}[T]} \\ &= e^{-(0.0187\ldots)\cdot 500} < 0.0000834. \end{aligned}$$

So the probability of getting 20% or more extra heads on 1000 coins is less than 1 in 10,000.

The bound rapidly becomes much smaller as the number of coins increases, because the expected number of heads appears in the exponent of the upper bound. For example, the probability of getting at least 20% extra heads on a million coins is at most

$$e^{-(0.0187\ldots)\cdot 500000} < e^{-9392},$$

which is an inconceivably small number.

Alternatively, the bound also becomes stronger for larger deviations. For example, suppose we’re interested in the odds of getting 30% or more extra heads in 1000 tosses, rather than 20%. In that case, $c = 1.3$ instead of 1.2. Consequently, the parameter $\beta(c)$ rises from 0.0187 to about 0.0410, which may not seem significant, but because $\beta(c)$ appears in the exponent of the upper bound, the final probability decreases from around 1 in 10,000 to about 1 in a billion!

20.5.4 Chernoff Bound for a Lottery Game

Pick-4 is a lottery game in which you pay \$1 to pick a 4-digit number between 0000 and 9999. If your number comes up in a random drawing, then you win \$5,000. Your chance of winning is 1 in 10,000. If 10 million people play, then the expected number of winners is 1000. When there are exactly 1000 winners, the lottery keeps \$5 million of the \$10 million paid for tickets. The lottery operator’s nightmare is that the number of winners is much greater—especially at the point where more than 2000 win and the lottery must pay out more than it received. What is the probability that will happen?

Let T_i be an indicator for the event that the ith player wins. Then $T = T_1 + \cdots + T_n$ is the total number of winners. If we assume[6] that the players’ picks and the winning number are random, independent and uniform, then the indicators T_i are independent, as required by the Chernoff bound.

[6] As we noted in Chapter 19, human choices are often not uniform and they can be highly dependent. For example, lots of people will pick an important date. The lottery folks should not get too much comfort from the analysis that follows, unless they assign random 4-digit numbers to each player.

Since 2000 winners would be twice the expected number, we choose $c = 2$, compute $\beta(c) = 0.386\ldots$, and plug these values into the Chernoff bound:

$$\begin{aligned}
\Pr[T \geq 2000] &= \Pr\big[T \geq 2\,\text{Ex}[T]\big] \\
&\leq e^{-k\,\text{Ex}[T]} = e^{-(0.386\ldots)\cdot 1000} \\
&< e^{-386}.
\end{aligned}$$

So there is almost no chance that the lottery operator pays out more than it took in. In fact, the number of winners won't even be 10% higher than expected very often. To prove that, let $c = 1.1$, compute $\beta(c) = 0.00484\ldots$, and plug in again:

$$\begin{aligned}
\Pr\big[T \geq 1.1\,\text{Ex}[T]\big] &\leq e^{-k\,\text{Ex}[T]} \\
&= e^{-(0.00484)\cdot 1000} < 0.01.
\end{aligned}$$

So the Pick-4 lottery may be exciting for the players, but the lottery operator has little doubt as to the outcome!

20.5.5 Randomized Load Balancing

Now let's return to Fussbook and its load balancing problem. Specifically, we need to determine a number m of servers that makes it very unlikely that any server is overloaded by being assigned more than 600 seconds of work in a given interval.

To begin, let's find the probability that the first server is overloaded. Letting T be the number of seconds of work assigned to the first server, this means we want an upper bound on $\Pr[T \geq 600]$. Let T_i be the number of seconds that the first server spends on the ith task: then T_i is zero if the task is assigned to another machine, and otherwise T_i is the length of the task. So $T = \sum_{i=1}^{n} T_i$ is the total number of seconds of work assigned to the first server, where $n = 24{,}000$.

The Chernoff bound is applicable only if the T_i are mutually independent and take on values in the range $[0, 1]$. The first condition is satisfied if we assume that assignment of a post to a server is independent of the time required to process the post. The second condition is satisfied because we know that no post takes more than 1 second to process; this is why we chose to measure work in seconds.

In all, there are 24,000 tasks, each with an expected length of 1/4 second. Since tasks are assigned to the m servers at random, the expected load on the first server is:

$$\begin{aligned}
\text{Ex}[T] &= \frac{24{,}000 \text{ tasks} \cdot 1/4 \text{ second per task}}{m \text{ servers}} \\
&= 6000/m \text{ seconds}. \qquad (20.24)
\end{aligned}$$

So if there are fewer than 10 servers, then the expected load on the first server is greater than its capacity, and we can expect it to be overloaded. If there are exactly 10 servers, then the server is expected to run for $6000/10 = 600$ seconds, which is 100% of its capacity.

Now we can use the Chernoff bound based on the number of servers to bound the probability that the first server is overloaded. We have from (20.24)

$$600 = c \operatorname{Ex}[T] \qquad \text{where } c ::= m/10,$$

so by the Chernoff bound

$$\Pr[T \geq 600] = \Pr[T \geq c \operatorname{Ex}[T]] \leq e^{-(c \ln(c) - c + 1) \cdot 6000/m},$$

The probability that *some* server is overloaded is at most m times the probability that the first server is overloaded, by the Union Bound in Section 17.5.2. So

$$\begin{aligned}
\Pr[\text{some server is overloaded}] &\leq \sum_{i=1}^{m} \Pr[\text{server } i \text{ is overloaded}] \\
&= m \Pr[\text{the first server is overloaded}] \\
&\leq m e^{-(c \ln(c) - c + 1) \cdot 6000/m},
\end{aligned}$$

where $c = m/10$. Some values of this upper bound are tabulated below:

$$\begin{array}{rcll}
m & = & 11: & 0.784\ldots \\
m & = & 12: & 0.000999\ldots \\
m & = & 13: & 0.0000000760\ldots.
\end{array}$$

These values suggest that a system with $m = 11$ machines might suffer immediate overload, $m = 12$ machines could fail in a few days, but $m = 13$ should be fine for a century or two!

20.5.6 Proof of the Chernoff Bound

The proof of the Chernoff bound is somewhat involved. In fact, *Chernoff himself* couldn't come up with it: his friend, Herman Rubin, showed him the argument. Thinking the bound not very significant, Chernoff did not credit Rubin in print. He felt pretty bad when it became famous![7]

[7] See "A Conversation with Herman Chernoff," *Statistical Science* 1996, Vol. 11, No. 4, pp 335–350.

Proof. (of Theorem 20.5.1)

For clarity, we'll go through the proof "top down." That is, we'll use facts that are proved immediately afterward.

The key step is to exponentiate both sides of the inequality $T \geq c\,\text{Ex}[T]$ and then apply the Markov bound:

$$\begin{aligned}
\Pr[T \geq c\,\text{Ex}[T]] &= \Pr[c^T \geq c^{c\,\text{Ex}[T]}] \\
&\leq \frac{\text{Ex}[c^T]}{c^{c\,\text{Ex}[T]}} && \text{(Markov Bound)} \\
&\leq \frac{e^{(c-1)\,\text{Ex}[T]}}{c^{c\,\text{Ex}[T]}} && \text{(Lemma 20.5.2 below)} \\
&= \frac{e^{(c-1)\,\text{Ex}[T]}}{e^{c\ln(c)\,\text{Ex}[T]}} = e^{-(c\ln(c)-c+1)\,\text{Ex}[T]}.
\end{aligned}$$

■

Algebra aside, there is a brilliant idea in this proof: in this context, exponentiating somehow supercharges the Markov bound. This is not true in general! One unfortunate side-effect of this supercharging is that we have to bound some nasty expectations involving exponentials in order to complete the proof. This is done in the two lemmas below, where variables take on values as in Theorem 20.5.1.

Lemma 20.5.2.

$$\text{Ex}\left[c^T\right] \leq e^{(c-1)\,\text{Ex}[T]}.$$

Proof.

$$\begin{aligned}
\text{Ex}\left[c^T\right] &= \text{Ex}\left[c^{T_1+\cdots+T_n}\right] && \text{(def of } T\text{)} \\
&= \text{Ex}\left[c^{T_1}\cdots c^{T_n}\right] \\
&= \text{Ex}\left[c^{T_1}\right]\cdots\text{Ex}[c^{T_n}] && \text{(independent product Cor 19.5.7)} \\
&\leq e^{(c-1)\,\text{Ex}[T_1]}\cdots e^{(c-1)\,\text{Ex}[T_n]} && \text{(Lemma 20.5.3 below)} \\
&= e^{(c-1)(\text{Ex}[T_1]+\cdots+\text{Ex}[T_n])} \\
&= e^{(c-1)\,\text{Ex}[T_1+\cdots+T_n]} && \text{(linearity of Ex}[\cdot]\text{)} \\
&= e^{(c-1)\,\text{Ex}[T]}.
\end{aligned}$$

The third equality depends on the fact that functions of independent variables are also independent (see Lemma 19.2.2). ■

Lemma 20.5.3.

$$\text{Ex}[c^{T_i}] \le e^{(c-1)\,\text{Ex}[T_i]}$$

Proof. All summations below range over values v taken by the random variable T_i, which are all required to be in the interval $[0, 1]$.

$$\begin{aligned}
\text{Ex}[c^{T_i}] &= \sum c^v \Pr[T_i = v] && \text{(def of Ex}[\cdot]\text{)}\\
&\le \sum (1 + (c-1)v) \Pr[T_i = v] && \text{(convexity—see below)}\\
&= \sum \Pr[T_i = v] + (c-1)v \Pr[T_i = v] \\
&= \sum \Pr[T_i = v] + (c-1) \sum v \Pr[T_i = v] \\
&= 1 + (c-1)\,\text{Ex}[T_i] \\
&\le e^{(c-1)\,\text{Ex}[T_i]} && \text{(since } 1 + z \le e^z\text{).}
\end{aligned}$$

The second step relies on the inequality

$$c^v \le 1 + (c-1)v,$$

which holds for all v in $[0, 1]$ and $c \ge 1$. This follows from the general principle that a convex function, namely c^v, is less than the linear function $1 + (c-1)v$ between their points of intersection, namely $v = 0$ and 1. This inequality is why the variables T_i are restricted to the real interval $[0, 1]$. ■

20.5.7 Comparing the Bounds

Suppose that we have a collection of mutually independent events $A_1, A_2, \ldots, A_n$, and we want to know how many of the events are likely to occur.

Let T_i be the indicator random variable for A_i and define

$$p_i = \Pr[T_i = 1] = \Pr[A_i]$$

for $1 \le i \le n$. Define

$$T = T_1 + T_2 + \cdots + T_n$$

to be the number of events that occur.

We know from Linearity of Expectation that

$$\begin{aligned}
\text{Ex}[T] &= \text{Ex}[T_1] + \text{Ex}[T_2] + \cdots + \text{Ex}[T_n] \\
&= \sum_{i=1}^{n} p_i .
\end{aligned}$$

This is true even if the events are *not* independent.

By Theorem 20.3.8, we also know that

$$\begin{aligned}\mathrm{Var}[T] &= \mathrm{Var}[T_1] + \mathrm{Var}[T_2] + \cdots + \mathrm{Var}[T_n] \\ &= \sum_{i=1}^{n} p_i(1 - p_i),\end{aligned}$$

and thus that

$$\sigma_T = \sqrt{\sum_{i=1}^{n} p_i(1 - p_i)}.$$

This is true even if the events are only pairwise independent.

Markov's Theorem tells us that for any $c > 1$,

$$\Pr[T \geq c\,\mathrm{Ex}[T]] \leq \frac{1}{c}.$$

Chebyshev's Theorem gives us the stronger result that

$$\Pr[|T - \mathrm{Ex}[T]| \geq c\sigma_T] \leq \frac{1}{c^2}.$$

The Chernoff Bound gives us an even stronger result, namely, that for any $c > 0$,

$$\Pr[T - \mathrm{Ex}[T] \geq c\,\mathrm{Ex}[T]] \leq e^{-(c\ln(c) - c + 1)\,\mathrm{Ex}[T]}.$$

In this case, the probability of exceeding the mean by $c\,\mathrm{Ex}[T]$ decreases as an exponentially small function of the deviation.

By considering the random variable $n - T$, we can also use the Chernoff Bound to prove that the probability that T is much *lower* than $\mathrm{Ex}[T]$ is also exponentially small.

20.5.8 Murphy's Law

If the expectation of a random variable is much less than 1, then Markov's Theorem implies that there is only a small probability that the variable has a value of 1 or more. On the other hand, a result that we call *Murphy's Law*[8] says that if a random variable is an independent sum of 0–1-valued variables and has a large expectation, then there is a huge probability of getting a value of at least 1.

[8] This is in reference and deference to the famous saying that "If something can go wrong, it probably will."

Theorem 20.5.4 (Murphy's Law)**.** *Let $A_1, A_2, \ldots, A_n$ be mutually independent events. Let T_i be the indicator random variable for A_i and define*

$$T ::= T_1 + T_2 + \cdots + T_n$$

to be the number of events that occur. Then

$$\Pr[T = 0] \le e^{-\text{Ex}[T]}.$$

Proof.

$$\begin{aligned}
\Pr[T = 0] &= \Pr[\overline{A}_1 \cap \overline{A}_2 \cap \ldots \cap \overline{A}_n] && (T = 0 \text{ iff no } A_i \text{ occurs})\\
&= \prod_{i=1}^{n} \Pr[\overline{A}_i] && (\text{independence of } A_i)\\
&= \prod_{i=1}^{n} (1 - \Pr[A_i]) \\
&\le \prod_{i=1}^{n} e^{-\Pr[A_i]} && (1 - x \le e^{-x} \text{ for } 0 \le x \le 1)\\
&= e^{-\sum_{i=1}^{n} \Pr[A_i]} \\
&= e^{-\sum_{i=1}^{n} \text{Ex}[T_i]} && (\text{since } T_i \text{ is an indicator for } A_i)\\
&= e^{-\text{Ex}[T]} && (\text{linearity of expectation}) \quad \blacksquare
\end{aligned}$$

For example, given any set of mutually independent events, if you expect 10 of them to happen, then at least one of them will happen with probability at least $1 - e^{-10}$. The probability that none of them happen is at most $e^{-10} < 1/22000$.

So if there are a lot of independent things that can go wrong and their probabilities sum to a number much greater than 1, then Theorem 20.5.4 proves that some of them surely will go wrong.

This result can help to explain "coincidences," "miracles," and crazy events that seem to have been very unlikely to happen. Such events do happen, in part, because there are so many possible unlikely events that the sum of their probabilities is greater than one. For example, someone *does* win the lottery.

In fact, if there are 100,000 random tickets in Pick-4, Theorem 20.5.4 says that the probability that there is no winner is less than $e^{-10} < 1/22000$. More generally, there are literally millions of one-in-a-million possible events and so some of them will surely occur.