CHAPTER 1

AN INTRODUCTION TO PROBABILITY

PREDICTING THE FUTURE

The term Predicting the Future conjures up images of veiled women staring into hazy crystal balls, or bearded men with darting eyes passing their hands over cups of tea leaves, or something else equally humorously mysterious. We call these people Fortune Tellers and relegate their "professions" to the regime of carnival side-show entertainment, along with snake charmers and the like. For party entertainment we bring out a Ouija board; everyone sits around the board in a circle and watches the board extract its mysterious "energy" from our hands while it answers questions about things to come.

On the other hand, we all seem to have firm ideas about the future based on consistent patterns of events that we have observed. We are pretty sure that there will be a tomorrow and that our clocks will all run at the same rate tomorrow as they did today. If we look in the newspaper (or these days, on the Internet), we can find out what time the sun will rise and set tomorrow—and it would be very difficult to find someone willing to place a bet that this information is not accurate. Then again, whether or not you will meet the love of your life tomorrow is not something you expect to see accurately predicted in the newspaper.

We seem willing to classify predictions of future events into categories of the *knowable* and the *unknowable*. The latter category is left to carnival

Probably Not: Future Prediction Using Probability and Statistical Inference,
by Lawrence N. Dworsky.

fortune tellers to illuminate. The former category includes "predictions" of when you'll next need a haircut, how much weight you'll gain if you keep eating so much pizza, and so on. There does seem to be, however, an intermediate area of knowledge of the future. Nobody knows for certain when you're going to die. An insurance company, however, seems able to consult its mystical Actuarial Tables and decide how much to charge you for a life insurance policy. How can it do this if nobody knows when you're going to die? The answer seems to lie in the fact that if you study thousands of people similar in age, health, life style, and so on, to you, you would be able to calculate an average life span—and that if the insurance company sells enough insurance policies with rates based upon this average, in a financial sense this is "as good" as if the insurance company knows exactly when you are going to die. There is, therefore, a way to describe life expectancies in terms of the expected behavior of large groups of people in similar circumstances.

When predicting future events, you often find yourself in situations such as this where you know something about future trends but you do not know exactly what is going to happen. If you flip a coin, you know you'll get either heads or tails but you don't know which. If you flip 100 coins, or equivalently flip one coin 100 times, however, you'd expect to get approximately 50 heads and 50 tails.

If you roll a pair of dice, you know that you'll get some number between two and twelve, but you don't know which number you'll get. However, in the case of the roll of a pair of dice, you do know that, in some sense, it's more likely that you'll get six than that you'll get two.

When you buy a new light bulb, you may see written on the package "estimated lifetime 1500 hours." You know that this light bulb might last 1346 hours, 1211 hours, 1587 hours, 2094 hours, or any other number of hours. If the bulb turns out to last 1434 hours, you won't be surprised; but if it only lasts 100 hours, you'd probably switch to a different brand of light bulbs.

There is a hint that in each of these examples, even though you couldn't accurately predict the future, you could find some kind of pattern that teaches you something about the nature of the future. Finding these patterns, working with them, and learning what knowledge can and cannot be inferred from them is the subject matter of the study of probability and statistics.

I can separate our study into two classes of problems. The first of these classes is understanding the likelihood that something *might* occur. We need a rigorous definition of likelihood so that we can be consistent in our evaluations. With this definition in hand, I can look at problems such as "How likely is it that you can make money in a simple coin flipping game?" or "How likely is it that a certain medicine will do you more good than harm in alleviating some specific ailment?" I'll have to define and discuss *random events* and the patterns that these events fall into, called Probability Distribution Functions (PDFs). This study is the study of Probability.

The second class of problems involves understanding how well you really know something. I will only present quantifiable issues, not “Does she really love me?” and “Is this sculpture truly a work of art?”

The uncertainties in how well we really know something can come from various sources. Let’s return to the example of light bulb. Suppose you’re the manufacturer of these light bulbs. Due to variations in materials and manufacturing processes, no two light bulb filaments (the thin wires in these bulbs that get white hot and glow brightly) are identical. There are variations in the lifetime of your product that you need to understand. The easiest way to learn the variations in lifetime would be to run all your light bulbs until they burn out and then look at the numbers, but for obvious reasons this is not a good idea. If you could find the pattern by just burning out some (hopefully a small percentage) of the light bulbs, then you have the information you need both to truthfully advertise your product and to work on improving your manufacturing process.

Learning how to do this is the study of Statistics. I will assume that we are dealing with a *stationary random process*. In a stationary random process, if nothing causal changes, we can expect that the nature of the pattern of the data already in hand will be the same as the nature of the pattern of future events of this same situation, and we use *statistical inference* to predict the future. In the practical terms of our light bulb manufacturer example, I am saying that so long as we don’t change anything, the factory will turn out bulbs with the same distribution of lifetimes next week as it did last week. This assertion is one of the most important characteristics of animal intelligence, namely the ability to discern and predict based upon patterns. If you think that only people can establish a pattern from historical data and predict the future based upon it, just watch your dog run to the door the next time you pick up his leash.

This light bulb problem also exemplifies another issue that I will have to deal with. We want to know how long the light bulb we’re about to buy will last. We know that no two light bulbs are identical. We also realize that our knowledge is limited by the fact that we haven’t measured every light bulb made. We must learn to quantify how much of our ignorance comes from each of these factors and develop ways to express both our knowledge and our lack of knowledge.

RULE MAKING

As the human species evolved, we took command of our environment because of our ability to learn. We learn from experience. Learning from experience is the art/science of recognizing patterns and then generalizing these patterns to a *rule*. In other words, the pattern is the relevant raw data that we’ve collected. A rule is what we create from our analysis of the pattern that we use to predict the future. Part of the rule is either one or several preferred extrapolations and responses. Successful pattern recognition is, for example, seeing that seeds from certain plants, when planted at the right time of the year and given the right amount of water, will yield food; and that the seed from a given plant will always yield that same food. Dark, ominous looking clouds usually precede a fierce storm, and it's prudent to take cover when such clouds are seen. Also, leaves turning color and falling off the trees means that winter is coming, and preparations must be made so as to survive until the following spring.

If we notice that every time it doesn't rain for more than a week our vegetable plants die, we would generate a rule that if there is no rain for a week, we need to irrigate or otherwise somehow water the vegetable garden. Implicit in this is that somewhere a "hypothesis" or "model" is created. In this case our model is that plants need regular watering. When the data are fit to this model, we quantify the case that vegetable plants need water at least once a week, and then the appropriate watering rule may then be created.

An interesting conjecture is that much, if not all, of what we call *the arts* came about because our brains are so interested in seeing patterns that we take delight and often find beauty in well-designed original patterns. Our eyes look at paintings and sculptures, our ears listen to music, our brains process the language constructs of poetry and prose, and so on. In every case we are finding pleasure in studying patterns. Sometimes the patterns are clear, as in a Bach fugue. Sometimes the patterns are harder to recognize, as in a surrealistic Picasso painting. Sometimes we are playing a game looking for patterns that just might not be there—as in a Pollock painting. Perhaps this way of looking at things is sheer nonsense, but then how can you explain how a good book or a good symphony (or rap song if that's your style) or a good painting can grab your attention and in some sense please you? The arts don't seem to be necessary for the basic survival of our species, so why do we have them at all?

A subtle rustling in the brush near the water hole at dusk sometimes—but not always—means that a man-eating tiger is stalking you. It would be to your advantage to make a decision and take action. Even if you're not certain that there's really a tiger present, you should err on the cautious side and beat a hasty retreat; you won't get a second chance. This survival skill is a good example of our evolutionary tendency to look for patterns and to react as if these patterns are there, even when we are not really sure that they indeed are there. In formal terms, you don't have all the data, but you do have *anecdotal* information.

Our prehistoric ancestors lived a very provincial existence. Life spans were short; most people did not live more than about 30 years. They didn't get to see more than about 10,000 sunrises. People outside their own tribe (and possibly some nearby tribes) were hardly ever encountered, so that the average person never saw more than a few hundred other people over the course of a lifetime. Also, very few people (other than members of nomadic tribes) ever traveled more than about 50 miles from where they were born. There are clearly many more items that could be added to this list, but the point has probably been adequately made: Peoples' brains never needed to cope with situations where there were hundreds of thousands or millions of data points to reconcile.

In today's world, however, things are very different: A state lottery could sell a hundred million tickets every few months. There are about six billion (that's six thousand million) people on the earth. Many of us (at least in North America and Western Europe) have traveled thousands of miles from the place of our birth many times; even more of us have seen movies and TV shows depicting places and peoples all over the world. Due to the ease with which people move around, a disease epidemic is no longer a local issue. Also, because we are aware of the lives of so many people in so many places, we know about diseases that attack only one person in a hundred thousand and tragedies that occur just about anywhere. If there's a vicious murderer killing teenage girls in Boston, then parents in California, Saskatoon, and London hear about it on the evening news and worry about the safety of their daughters.

When dealing with unlikely events spread over large numbers of opportunities, your intuition can and does often lead you astray. Since you cannot easily comprehend millions of occurrences, or lack of occurrences, of some event, you tend to see patterns in a small numbers of examples—again the *anecdotal* approach. Even when patterns don't exist, you tend to invent them; you are using your "better safe than sorry" prehistoric evolved response. This could lead to the inability to correctly make many important decisions in your life: What medicines or treatments stand the best chance of curing your ailments? Which proffered medicines have been correctly shown to be useful, and which ones are simply quackery? Which environmental concerns are potentially real and which are simple coincidence? Which environmental concerns are no doubt real but probably so insignificant that it we can reasonably ignore them? Are "sure bets" on investments or gambling choices really worth anything? We need an organized methodology for examining a situation and coping with information, correctly extracting the pattern and the likelihood of an event happening or not happening to us, and also correctly "processing" a large set of data and concluding, when appropriate, that there really is or is not a pattern present.

In other words, we want to understand how to cope with a barrage of information. We need a way of measuring how sure we are of what we know, and when or if what we know is adequate to make some predictions about what's to come.

RANDOM EVENTS AND PROBABILITY

This is a good place to introduce the concepts of random events, random variables, and probability. These concepts will be wrung out in detail in later chapters, so for now let's just consider some casual definitions.

For our purposes an *event* is a particular occurrence of some sort out of a larger set of possible occurrences. Some examples are:

- Will it rain tomorrow? The full set of possible occurrences is the two events Yes—it will rain, and No—it won't rain.
- When you flip a coin, there are two possible events. The coin will either land head side up or tail side up (typically referred to as "heads" or "tails").
- When you roll one die, then there are six possible events, namely the six faces of the die that can land face up—that is, the numbers 1, 2, 3, 4, 5, and 6.
- When you play a quiz game where you must blindly choose "door A, door B, or door C" and there is a prize hiding behind only one of these doors, then there are three possible events: The prize is behind door A, it's behind door B, or it's behind door C.

Variable is a name for a number that can be assigned to an event. If the events themselves are numbers (e.g., the six faces of the die mentioned above), then the most reasonable thing to do is to simply assign the variable numbers to the event numbers. A variable representing the days of the year can take on values 1, 2, 3, . . . , all the way up to 365. Both of these examples are of variables that must be integers; that is, 4.56 is not an allowed value for either of them. There are, of course, cases where a variable can take on any value, including fractional values, over some range; for example, the possible amount of rain that fell in Chicago last week can be anything from 0 to 15 inches (I don't know if this is true or not, I just made it up for the example). Note that in this case 4.56, 11.237, or 0.444 are legitimate values for the variable to assume. An important distinction between the variable in this last example and the variables in the first two examples is that the former two variables only can take on a finite number of possibilities (6 in the first case, 365 in the second), whereas by allowing fractional values (equivalently, real number values), there are an infinite number of possibilities for the variable in the last example.

A random variable is a variable that can take on one of an allowed set of values (finite or infinite in number). The actual value selected is determined by a happening or happenings that are not only outside our control but also are outside of any recognized, quantifiable, control—but often do seem to follow some sort of pattern.

A random variable cannot take on any number, but instead must be chosen out of the set of possible occurrences of the situation at hand. For example, tossing a die and looking at the number that lands facing up will give us one of the variables {1, 2, 3, 4, 5, 6}, but never 7, 0, or 3.2.

The most common example of a simple random variable is the outcome of the flip of our coin. Let's assign the number –1 to a tail and +1 to a head. The flip of the coin must yield one of the two chosen values for the random variable, but we seem to have no way of predicting which value it will yield for a specific flip.

Is the result of the flip of a coin truly unpredictable? Theoretically, no: If you carefully analyzed the weight and shape of the coin and then tracked the exact motion of the flipper's wrist and fingers, along with the air currents present and the nature of the surface that the coin lands on, you would see that the flipping of a coin is a totally predictable event. However, since it is so difficult to track all these subtle factors carefully enough in normal circumstances and these factors are extremely difficult to duplicate from flip to flip, the outcome of a coin flip can reasonably be considered to be a random event. Furthermore, you can easily list all the possible values of the random variable assigned to the outcome of the coin flip (–1 or 1); and if you believe that the coin flip is fair, you conclude that either result is equally likely. This latter situation isn't always the case.

If you roll two dice and define the random variable as the sum of the numbers you get from each die, then this random variable can take on any value from 2 to 12. All of the possible results, however, are no longer equally likely. This assertion can be understood by looking at every possible result as shown in Table 1.1.

As may be seen from the table, there is only one way that the random variable can take on the value 2: Both dice have to land with a 1 face up. However, there are three ways that the random variable can take on the value 4: One way is for the first die to land with a 1 face up while the second die lands with a three face up. To avoid writing this out over and over again, I'll call this case {1, 3}. By searching through the table, we see that the random variable value of 4 can be obtained by the dice combinations {1, 3}, {2, 2}, and {3, 1}.

I'll create a second table (Table 1.2) that tabulates the values of the random variable and the number of ways that each value can result from the rolling of a pair of dice:

The numbers in the right-hand column add up to 36. This is just a restatement of the fact that there are 36 possible outcomes possible when rolling a pair of dice.

Define the *probability* of a random event as the number of ways that that event can occur, divided by the number of all possible events. Adding a third column to the table to show the probabilities, I get Table 1.3.

For example, if you want to know the probability that the sum of the numbers on the two dice will be 5, the second column of this table tells us that there are four ways to get 5. Looking back at the first table, you can see that this comes about from the possible combinations {1, 4}, {2, 3}, {3, 2} and {4, 1}. The probability of rolling two dice and getting a (total) of 5 is therefore 4/36,

TABLE 1.1. All the Possible Results of Rolling a Pair of Dice

First Die Result	Second Die Result	Random Variable Value = Sum of First & Second Results	First Die Result	Second Die Result	Random Variable Value = Sum of First & Second Results
1	1	2	4	1	5
1	2	3	4	2	6
1	3	4	4	3	7
1	4	5	4	4	8
1	5	6	4	5	9
1	6	7	4	6	10
2	1	3	5	1	6
2	2	4	5	2	7
2	3	5	5	3	8
2	4	6	5	4	9
2	5	7	5	5	10
2	6	8	5	6	11
3	1	4	6	1	7
3	2	5	6	2	8
3	3	6	6	3	9
3	4	7	6	4	10
3	5	8	6	5	11
3	6	9	6	6	12

TABLE 1.2. Results of Rolling a Pair of Dice Grouped by Results

Value of Random Variable	Number of Ways of Obtaining this Value
2	1
3	2
4	3
5	4
6	5
7	6
8	5
9	4
10	3
11	2
12	1

TABLE 1.3. Same as Table 1.2 but also Showing Probability of Results

Value of Random Variable	Number of Ways of Obtaining this Result	Probability of Getting this Result
2	1	1/36 = 0.028
3	2	2/36 = 0.056
4	3	3/36 = 0.083
5	4	4/36 = 0.111
6	5	5/36 = 0.139
7	6	6/36 = 0.167
8	5	5/36 = 0.139
9	4	4/36 = 0.111
10	3	3/36 = 0.083
11	2	2/36 = 0.056
12	1	1/36 = 0.028

sometimes called “4 chances out of 36.” 4/36 is of course the same as 2/18 and 1/9 and the decimal equivalent, 0.111.[1]

If you add up all of the numbers in the new rightmost column, you’ll get exactly 1. This will always be the case, because it is the sum of the probabilities of all possible events. This is the “certain event” and it must happen; that is, it has a probability of 1 (or 100%). This certain event will be that, when you toss a pair of dice, the resulting number—the sum of the number of dots on the two faces that land face up—again must be some number between 2 and 12.

Sometimes it will be easier to calculate the probability of something we’re interested in *not* happening than to calculate the probability of it happening. In this case since we know that the probability of our event either happening or not happening must be 1, then the probability of the event happening is simply 1—the probability of the event not happening.

From Table 1.3 you can also calculate combinations of these probabilities. For example, the probability of getting a sum of *at least 10* is just the probability of getting 10 + the probability of getting 11 + the probability of getting 12, = 0.083 + 0.056 + 0.028 = 0.167. Going forward, just for convenience, we’ll use the shorthand notation Prob(12) to mean “the probability of getting 12,” and we’ll leave some things to the context; that is, when rolling a pair of dice, we’ll assume that we’re always interested in the sum of the two numbers facing up, and we’ll just refer to the number.

Exactly what the probability of an event occurring really means is a very difficult and subtle issue. Let’s leave this for later on, and just work with the

[1] Many fractions, such as 1/9, 1/3, and 1/6, do not have exact decimal representations that can be expressed in a finite number of digits. 1/19, for example, is 0.111111111 . . . , with the 1’s going on forever. Saying that the decimal equivalent of 1/9 is 0.111 is therefore an approximation. Knowing how many digits are necessary to achieve a satisfactory approximation is context-dependent—there is no easy rule.

intuitive “If you roll a pair of dice very many times, about 1/36 of the time the random variable will be 2, about 2/36 of the time it will be 3, and so on.”

An alternative way of discussing probabilities that is popular at horse races, among other places, is called *the odds* of something happening. Odds is just another way of stating things. If the probability of an event is 1/36, then we say that the odds of the event happening is 1 to 35 (usually written as the ratio 1:35). If the probability is 6/36, then the odds are 6:30 or 1:5, and so on. As you can see, while the *probability* is the number of ways that a given event can occur divided by the total number of possible events, the *odds* is just the ratio of the number of ways that a given event can occur to the number of ways that it can’t occur. It’s just another way of expressing the same calculation; neither system tells you any more or less than the other.

In the simple coin flip game, the probability of winning equals the probability of losing, = 0.5. The odds in this case is simply 1:1, often called *even odds*. Another *term of art* is the case when your probability of winning is something like 1:1000. It’s very unlikely that you’ll win; these are called *long odds*.

Something you’ve probably noticed by now is that I tend to jump back and forth between fractions (such as 1/4) and their decimal equivalents (1/4 = 0.25). Mathematically, it doesn’t matter which I use. I tend to make my choice based on context: When I want to emphasize the origins of the numerator and denominator (such as 1 chance out of 4), I’ll usually use the fraction, but when I just need to show a number that’s either the result of a calculation or that’s needed for further calculations, I’ll usually use the decimal. I hope this style pleases you rather than irritates you; the important point is that insofar as the mathematics is concerned, both the fraction and the decimal are equivalent.

You now have the definitions required to look at a few examples. I’ll start with some very simple examples and work up to some fairly involved examples. Hopefully, each of these examples will illustrate an aspect of the issues involved in organizing some probabilistic data and drawing the correct conclusion. Examples of statistical inference will be left for later chapters.

THE LOTTERY {VERY IMPROBABLE EVENTS AND VERY LARGE DATA SETS}

Suppose you were told that there is a probability of 1 in 200 million (that’s 0.000000005 as a decimal) of you getting hit by a car and being seriously injured or even killed if you leave your house today. Should you worry about this and cancel your plans for the day? Unless you really don’t have a very firm grip on reality, the answer is clearly *no*. There are probabilities that the next meal you eat will poison you, that the next time you take a walk it will start storming and you’ll be hit by lightening, that you’ll trip on your way to the bathroom and split your skull on something while falling, that an airplane will fall out of the sky and crash through your roof, and so on. Just knowing that you and your acquaintances typically do make it through the day is anecdotal evidence that the sum of these probabilities can't be a very large number. Looking at your city's accidental death rate as a fraction of the total population gives you a pretty realistic estimate of the sum of these probabilities. If you let your plans for your life be compromised by every extremely small probability of something going wrong, then you will be totally paralyzed.[2] One in two hundred million, when it's the probability of something bad happening to you, might as well be zero.

Now what about the same probability of something good happening to you? Let's say you have a lottery ticket, along with 199,999,999 other people, and one of you is going to win the grand prize. Should you quit your job and order a new car based on your chance of winning?

The way to arrive at an answer to this question is to calculate a number called the expected value (of your winnings). I'll define expected value carefully in the next chapter, but for now let me just use the intuitive "What should I expect to win?" There are 4 numbers I need in order to perform the calculation.

First, I need the probability of winning. In this case it's 1 in 200 million, or 0.000000005. Next, I need the probability of losing. Since the probability of losing plus the probability of winning must equal 1, the probability of losing must be $1-0.000000005=.999999995$.

I also need the amount of money you will make if you win. If you buy a lottery ticket for \$1 and you will get \$50,000,000 if you win, this is \$50,000,000 − \$1 = \$49,999,999.

Lastly, I need the amount of money you will lose if you don't win. This is the dollar you spent to buy the lottery ticket. Let's adopt the sign convention that winnings are a positive number but losses are a negative number. The amount you'll lose is therefore –\$1.

In order to calculate the expected value of your winnings, I add up the product of each of the possible money transfers (winning and losing) multiplied by the probability of this event. Gathering together the numbers from above, we obtain

$$\begin{aligned}\text{Expected value} &= (0.000000005)(\$49{,}999{,}999)-(.999999995)(\$1)\\ &\approx (0.000000005)(\$50{,}000{,}000)-(1)(\$1)=\$0.25-\$1.00=-\$0.75\end{aligned}$$

I have just introduced the symbol "≈", which means "not exactly, but a good enough approximation that the difference is irrelevant." "Irrelevant," of course, depends on the context of the situation. In this example, I'm saying

[2] In 1976, when the U.S. Skylab satellite fell from the sky, there were companies selling Skylab insurance—coverage in case you or your home got hit. If you consider the probability of this happening as approximately the size of the satellite divided by the surface area of the earth, you'll see why many fortunes have been made based on the truism that "there's a sucker born every minute."

that (0.000000005)($49,999,999) = $0.249999995 is close enough to $0.25 that when we compare it to $1.00 we never notice the approximation.

The expected value of your winnings is a negative number—that is, you should expect to lose money. What the expected value is actually telling you is that if you had bought *all* of the lottery tickets, so that you had to be the winner, you would still lose 75 cents on every dollar you spent. It's no wonder that people who routinely calculate the value of investments and gambling games often refer to lotteries as a "Tax on Stupidity."

What I seem to be saying so far is that events with extremely low probabilities simply don't happen. If we're waiting for you to win the lottery, then this is a pretty reasonable conclusion. However, the day after the lottery drawing there will be an article in the newspaper about the lottery, along with a picture of a very happy person holding up a winning lottery ticket. This person just won 50 million dollars!

Am I drawing two different conclusions from the same set of data? Am I saying both that nobody wins the lottery and that somebody always wins the lottery? The answer is that there is no contradiction, we just have to be very careful how we say what we say. Let me construct an example. Suppose the state has a lottery with the probability of any one ticket winning = 0.000000005 and the state sells 200 million tickets, which include every possible choice of numbers. It's an absolute certainty that *somebody* will win (we'll ignore the possibility that the winning ticket got accidentally tossed into the garbage). This does not at all contradict the statement that it's "pretty darned near" certain that *you* won't win.

What we are struggling with here is the headache of dealing with a very improbable event juxtaposed on a situation where there are a huge number of opportunities for the event to happen. It's perfectly reasonable to be assured that something will never happen to you while you know that it will happen to somebody. Rare diseases are an example of this phenomenon. You shouldn't spend much time worrying about a disease that randomly afflicts one person in, say, 10 million, every year. But in the United States alone there will be about 30 cases of this disease reported every year, and from a Public Health point of view, somebody should be paying attention to it.

A similar situation arises when looking at the probability of an electrical appliance left plugged in on your countertop starting a fire. Let's say that this probability is 1 in 30,000 per person.[3] Should you meticulously unplug all your countertop kitchen appliances when you're not using them? Based on the above probability, the answer is "don't bother." However, what if you're the senior fire department safety officer for New York City, a city with about 8 million residents? I'll assume an average of about 4 people per residence. If

[3] The U.S. Fire Administration's number is about 23,000 appliance related electrical fires per person. I rounded this up to 30,000 to make a convenient comparison to a population of about 300 million.

nobody unplugs their appliances, then you're looking at about 20 unnecessary fires every year, possible loss of life, and certain destruction of property. You are certainly going to tell people to unplug their appliances. This is a situation where the mathematics might not lead to the right answer, assuming that there is a right answer. You'll have to draw your own conclusion here.

One last note about state lotteries. In this example, the state took in $200 million and gave out $50 million. In principle, this is why state lotteries are held. The state makes money that it uses for education or for health care programs for the needy, and so on. From this perspective, buying a lottery ticket is both a social donation and a bit of entertainment for you—that's not a bad deal for a dollar or two. On the other hand, as an investment this not only has an absurdly low chance of paying off, but since the expected value of the payoff is negative, this is not what's called a *Fair Game*. From an investment point of view, this is a very poor place to put your money.

COIN FLIPPING {FAIR GAMES, LOOKING BACKWARDS FOR INSIGHT}

Let's set up a really simple coin flipping game between you and a friend. One of you flips a coin. If the result is heads, you collect a dollar from your friend; if the result is tails, you pay a dollar to your friend. Assuming that the coin is fair (not weighted, etc.), there is an equal probability of getting either a head or a tail. Since the sum of all the probabilities must equal one, then the probability of getting a head and the probability of getting a tail must be equal to 0.5 (one-half).

Letting +$1 be the result of winning (you get to be a dollar richer) and letting –$1 be the result of losing (you get to be a dollar poorer), then the expected value of your return is

$$E = (0.5)(+1\$) + (0.5)(-1\$) = 0$$

This is what I'm calling a fair game. Since I am defining positive $ values as winning (money coming to you) and defining negative $ values as losing (money leaving you), then if your winnings and your losings exactly balance out, the algebraic sum is zero. Nobody involved (in this case there is just you and your friend) should *expect* to win, or lose, money. This last sentence seems at first to be very simple. Examining the nuances of what is meant by *expect* thoroughly, however, will take a significant portion of this book.

While Expected Value is a mathematical term that is carefully defined, the definition is not quite the same as the common, conversational, use of the term *expected* or *expectation*. Also, as a few examples will show, in most cases it is much less likely that you will come away from this game with the exact amount of money that you went into the game with than that you either win or lose some money.

The simplest example of this last claim is a game where you flip the coin once and then stop playing. In this case you must either win or lose a dollar—there are no other choices. For that matter, if the game is set up so that there will be any odd number of coin flips, it is impossible for you to come away neither winning nor losing. In these cases it is impossible to win the *expected* amount of winning—clearly a distinction between the mathematical definition of Expected Value and its common usage.

Now let's look at some games where there are an even number of coin flips. In these cases it certainly is possible to end up with a zero sum. The simplest example is the two-flip game. If you flip first a head and then a tail (or vice versa), you come away with a zero sum.

As a brief aside, let's introduce some mathematical notation that will make things easier as we proceed. If we let the letter n refer to the number of coin flips, then our two-flip game can be referred to as an $n=2$ game. The choice of the letter n was completely arbitrary. No new ideas or calculations have just been presented. This is simply a convenient way to talk about things. At this point there is very little to be gained by doing this, but the approach will prove to be very powerful and useful as things get more complicated.

Returning to the examination of coin flipping, let's examine the $n=2$ game in detail. This is easy to do because there are only four possible scenarios. I'll show both the results of the flip (heads or tails) and the algebraic value of the random variable associated with these results, just for clarity in Table 1.4.

As the table shows, there are two opportunities for a zero sum: one opportunity for a positive sum and one opportunity for a negative sum. The probability that you will win is therefore 0.25, the probability that you'll lose is 0.25, and the probability that you'll break even is 0.5. In other words, it's equally likely that there will be a zero sum and that there will be a nonzero sum.

Now let's look at a game with a larger value of n. What about, say, $n=10$? How many table entries must there be? For each flip of the coin, there are two possibilities. Therefore, for $n=2$ there are 4 possibilities, for $n=3$ there are $(2)(2)(2)=8$ possibilities, for $n=4$ there are $(2)(2)(2)(2)=16$ possibilities, and so on. Adding to our mathematical notation toolbox, the expression 2^n means 2 multiplied by itself n times. In other words, $2^3=8$, $2^4=16$, and so on. We can therefore say that "There are 2^n possibilities for an n-flip coin flip game."

A 10-flip game would have 2^{10} possible outcomes. Working this out, we obtain

TABLE 1.4. All the Possible Results of an $n=2$ Coin Flip Game

First Flip	First Flip Variable	Second Flip	Second Flip Variable	Sum
Head	+1	Tail	–1	0
Head	+1	Head	+1	2
Tail	–1	Tail	–1	–2
Tail	–1	Head	+1	0

$$2^{10} = 1024$$

While there is no reason why I couldn't create the list and examine every possible outcome, it certainly does not look like doing this would be fun. Let's set our goals a little more humbly, I'll look at $n = 3$ and $n = 4$ games. Adding to our notational soup mix, let the letter k refer to a particular flip; that is, $k = 2$ refers to the second flip, $k = 3$ refers to the third flip, and so on. Clearly, in an nth-order game, k can take on all of the (integer) values from 1 to n. In algebraic notation, this is written as

$$1 \le k \le n$$

The symbol $\le$ means "less than or equal to," so the above expression is read as "1 is equal to or less than k, and also k is equal to or less than n." It may not seem that way now, but this is actually a very convenient way to express things.

Using our new notational ability, Table 1.5 shows the $n = 3$ game.

As expected, there is absolutely no way to play an $n = 3$ game and come out even; you have to either win or lose some amount. However, since each of the outcomes (rows) in the above table is equally likely, and there are 8 of them, the probability of each outcome is exactly 1/8, so that the expected value of the your return is

$$\frac{-3}{8} + \frac{-2}{8} + \frac{-1}{8} + \frac{+1}{8} + \frac{-1}{8} + \frac{+1}{8} + \frac{+2}{8} + \frac{+3}{8} = \frac{0}{8} = 0$$

The expected value is zero, so this is indeed a fair game. Worth noting again in this case is the fact that the expected value might be a value that you can never realize. You can play $n = 3$ games all night, and you will never come away with a 0 (no loss or no gain) result from one of these games. What you

TABLE 1.5. All the Possible Results of an $n = 3$ Coin Flip Game

$K = 1$	$k = 2$	$k = 3$	Sum
–1	–1	–1	–3
–1	–1	+1	–2
–1	+1	–1	–1
–1	+1	+1	+1
+1	–1	–1	–1
+1	–1	+1	+1
+1	+1	–1	+2
+1	+1	+1	+3

might expect, however, is that after a night of $n = 3$ games, the average of your results would be close to zero.

In this simple coin flip situation, a hundred $n = 3$ games is exactly the same as an $n = 300$ game. We can therefore extend our logic: For any number of $n =$ anything games, if the total number of coin flips is odd, you can still never walk away with a zero sum. You might, however, come close.

Let's look at an $n = 4$ game and see what happens (Table 1.6):

In an $n = 4$ game, there are $2^4 = 16$ possible outcomes, all listed in Table 1.6. Looking through the column of sums, we see that there are 6 possible ways to get 0. In other words, the probability of neither winning nor losing is 6/16 = .375. This is lower than the probability of neither winning nor losing was for an $n = 2$ game. Is this a pattern?

Table 1.7 shows that the probability of getting a zero sum out of even order games for $2 \leq n \leq 20$.

There is indeed a trend: As n gets larger, the probability of getting a zero sum gets lower. In other words, the more times you flip the coin, the less likely it is that you will get exactly the same number of heads and tails.

TABLE 1.6. All the Possible Results of an *n* = 4 Coin Flip Game

$k = 1$	$k = 2$	$k = 3$	$k = 4$	Sum	$k = 1$	$k = 2$	$k = 3$	$k = 4$	Sum
–1	–1	–1	–1	–4	1	–1	–1	–1	–2
–1	–1	–1	1	–2	1	–1	–1	1	0
–1	–1	1	–1	–2	1	–1	1	–1	0
–1	–1	1	1	0	1	–1	1	1	2
–1	1	–1	–1	–2	1	1	–1	–1	0
–1	1	–1	1	0	1	1	–1	1	2
–1	1	1	–1	0	1	1	1	–1	2
–1	1	1	1	2	1	1	1	1	4

TABLE 1.7. Probabilities of 0 Sum for *n* = 2 to *n* = 20 Coin Flip Games

n	Probability of 0 Sum
2	0.500
4	0.375
6	0.313
8	0.273
10	0.246
12	0.226
14	0.209
16	0.196
18	0.185
20	0.176

This leads to some very thought-provoking discussions of just what a probability means, what you can expect, and how sure you can be of what you expect. This, in turn, leads to a very basic question of just what it means to be sure about something, or in other words, just how confident you are about something happening. In the Chapter 3 I'll define a "confidence factor" by which we can gauge just how sure we are about something.

Returning to the simple coin flip, what if I want the probability of getting 5 heads in a row? This is identically the probability of flipping 5 coins and getting 5 heads, because the results of coin flips, be they with multiple coins or with the same coin over and over again, are independent of each other. You could get the answer by writing a list such as the tables I've been presenting for the $n=5$ case or you could just note that for independent events, all you have to do is to multiply the individual probabilities together. In other words, the probability of getting 5 heads in a row is just $(1/2)^5 = 1/32 \approx 0.033$.

1/32 is a low probability, but not so low as to astound you if it happens. What if it indeed just did happen? You flipped a coin and got 5 heads in a row. What's the probability of getting a head if you now flip the coin again? Assuming, of course, that the coin isn't weighted or corrupted in any other manner (i.e., that the flips are indeed fair), the probability of a head on this flip (and for any subsequent flip) is still just 1/2. Putting it simply, a coin has no memory.

Reiterating the point above, flipping one coin 6 times is statistically identical to flipping six different coins once each and then examining the results. It doesn't matter whether you flip the six coins one at a time or if you toss them all up into the air and let them fall onto the table. The six coins are independent of each other: They do not "know" or "remember" anything about either their own past performance or the performance of any other coin. When you look at it this way, it's pretty clear that the flip of the sixth coin has nothing to do with the flips of the first five coins. For that matter, if you tossed all six coins into the air at once, you couldn't even say which coin is the "sixth coin."

The above arguments are a simple case of another interesting discussion: When does knowledge of past results tell you something about future results? In the above example, it doesn't tell you anything at all. Later in this chapter I will show an example where this isn't the case.[4]

Returning to the coin flipping game, remember that the expected value of return of an $n =$ anything coin flip is always zero, as has been illustrated in several examples. If you were to flip a coin 10 times, the most likely single result would be an equal number of heads and tails, even though it's not a very likely event (remember, in this case we're counting the number of ways

[4] People who don't understand this point will tell you that if you flipped a coin 100 times and got 100 heads, you should bet on a tail for the 101st flip because "it's due." I'd be more inclined to bet on a 101st head, because after the first 100 heads I'd be pretty sure that I wasn't dealing with a fair coin.

of getting 5 heads and 5 tails out of $2^{10}=1024$ possible configurations). The distinction between *most* likely and yet *not very* likely (or the equivalent, very unlikely) eludes many people, so let's consider another example.

Suppose I have a giant roulette wheel with 1000 slots for the ball to land in. I'll number the slots 1 to 999 consecutively, and then number the thousandth slot 500. This means I have exactly one slot for each number between 1 and 999, except for the number 500, for which I have two slots. When I spin this roulette wheel and watch for the ball to settle in a slot, I see that there are two opportunities for the ball to settle in a slot numbered 500, but only one opportunity for the ball to settle at any other number. In other words, the probability of the ball landing at 500 is twice the probability of the ball landing at any other number. 500 is clearly the most likely result. The probability of the ball landing in a 500 slot is 2 out of 1000, or 0.002. The probability of the ball *not* landing in a 500 slot—that is, the probability of landing *anywhere but* in a 500 slot—is 998 out of 1000, or 0.998. It is, therefore, very unlikely that the ball will land in a 500 slot. Now let's combine both of the above observations into the same sentence: The most likely slot the ball will land in will be numbered 500, but it is very unlikely that the ball will land in a slot numbered 500 as compared to some other number.

Returning to the coin flipping example, no matter how many (even number of) times you flip a coin, the most likely result is that you'll get an equal number of heads and tails. The more times you flip the coin, however, the less likely this result will be. This same idea will be presented in the chapter on random walks.

A variation on the above example is the picking of a number for a lottery. If you were to pick, say, 12345, or 22222, you would be criticized by the "experts": "You never see a winning number with a regular pattern—it's always something like 13557 or 25738 or. . . . " This last statement is correct. It is correct because of all the nearly one million five-digit numbers that can be picked, very few of them have simple, recognizable digit patterns. It is therefore most likely that the winning number will *not* have a recognizable pattern. However, the five lottery balls have no memory or awareness of each other. They would not "know" if they presented a recognizable pattern. Any five-digit number is equally likely. The difference between the recognizable patterns and other patterns is only in the eyes of the beholder, I can't even imagine how I'd define a "recognizable" pattern except maybe by the majority vote of a room full of people.

A corollary to this is of course that there's no reason not to pick the very number that won last week. It's highly unlikely that this number will win again just because there are so many numbers to choose from, but it is just as likely that this number will win as it is that any other number will win.

Moving on, what if you have just flipped a coin five times, got five heads, and now want to flip the coin ten times more? The expectation value looking forward is still zero. But, having just won the game five times, you have five dollars more in your pocket than you started with. Therefore, the most likely scenario is that you will end up with five dollars in your pocket! This property will be covered in more detail in the chapter on gambling games (Chapter 8). Generalizing this conclusion, I can say that if you are going to spend the evening flipping coins, your most likely status at the finish is just your status at the time you think about it. If you start off lucky (i.e., have net winnings early on), then you'll probably end up winning a bit, and vice versa. There really is such a thing as "being on a winning streak." However, this observation can only be correctly made after the fact. If you were lucky and got more heads than tails (or vice versa, if that's the side you're on), then you were indeed on a winning streak. The perception that being on a winning streak so far will influence the coin's future results is of course total nonsense. You might win a few times in a row, you might even win most of the time over the course of the evening, but each flip is still independent of all the others.

There is an argument for saying that if you have been on a winning streak, it's more likely that you'll end the evening ahead (i.e., with net winnings rather than losses) than if you haven't been on a winning streak. That argument is that if you have been on a winning streak, you have a lot more money in your pocket than you would have if you had been on a losing streak. You are therefore in a better position to withstand a few (more) losses without being wiped out and having to quit playing, and therefore your odds of winning for the evening have been increased. This argument has nothing to do with the probabilities of an individual win (coin flip, roulette wheel, poker hand, whatever). If you are just playing for a score on a piece of paper and cannot be "wiped out," this argument is worthless.

THE COIN FLIP STRATEGY THAT CAN'T LOSE

Assume that you want to earn $1 a day (scale this to $10, or $100, or anything you wish—the discussion is clearest when working with a starting value of $1). Let's build a strategy for playing a clever coin flipping game:

1. Bet $1 on the results of a coin flip.
2. If you win the first coin flip, you've earned your $1 for the day. Go home.
3. If you lose the first coin flip, you've lost $1. Bet $2 and try again.
4. If you win the second coin flip, then you've recovered your lost $1 and won $1. Go home.
5. If you lose the second coin flip, then you've now lost $3. Bet $4 and try again.
6. If you win the third coin flip, then you've recovered your lost $7 and won $1. Go home.
7. If you lose the third coin flip, then you've now lost $7. Bet $8 and try again.
8. And so on, until you win.

This scheme seems unbeatable. If you keep flipping a coin, sooner or later you have to get a head, and you've won for the day. What could possibly go wrong?

The scheme could be analyzed in detail from a number of different perspectives. The Achilles' heel is that you need a very big wallet. For example, to cover three losses and still have the money to place your fourth bet, you need to start with $\$1+\$2+\$4+\$8=\$15$. In general, to be able to place n bets you need to start out with 2^n-1 dollars.

If you start the evening with 1 million dollars, it's pretty certain that you will be able to go home with $1 million + $1. You have enough money to cover a lot of tails until you get your head. On the other hand, if you show up with only $1, then the probability of you going home with your starting $1 plus your winnings of $1 is only 50%. Putting this last case slightly differently, the probability of you doubling your money before getting wiped out is 50%. Without showing the details now, let me just say that it turns out that no matter now much money you start with, the probability of you doubling your money before you get wiped out is at best 50%, less if you have to flip the coin many times. If you're "earning" only a dollar a day, then you need to come back the number of days equal to the number of dollars you're starting with. You can save yourself a lot of time, however, by just betting all of your money on the first coin flip—a very simple game with a 50% probability of doubling your money and a 50% probability of being wiped out. Here again, we have been reconciling an unlikely event (not getting a head after many flips of a coin) with a large number of opportunities for the event to happen (many flips of a coin). We should note here that it's not possible to "fool Mother Nature." If you start out with a million dollars and plan to visit 10,000 gambling houses each night, hoping to win only $1 at each house, then the probabilities start catching up with you and you no longer have a sure thing going.

THE PRIZE BEHIND THE DOOR {LOOKING BACKWARDS FOR INSIGHT, AGAIN}

This example is subtle, and the answer is often incorrectly guessed by people who should know better. There are still ongoing debates about this puzzle on various Probability-Puzzle websites because the correct answer seems to be so counterintuitive to some people that they just won't accept the analysis. It is known by many names, perhaps most commonly the "Monty Hall" problem, named after the host of a TV game show.

You are a participant in a TV game show. There are three doors (let's call them doors A, B, and C). Behind one of these doors is a substantial prize, behind the other two doors is nothing. You have to take a guess. So far this is very straightforward: Your probability of guessing correctly and winning the prize is exactly 1/3.

You take (and announce) your guess. Before the three doors are opened to reveal the location of the prize, the game show host goes to one of the two doors that you *didn't* choose, opens it, and shows you that the prize is *not* behind this door. The prize, therefore, must either be behind the unopened door that you chose or behind the unopened door that you did not choose. You are now given the option of staying with your original choice or switching to the unopened door that you did not choose. What should you do?

Almost everybody's first response to this puzzle is to shrug—after all, there are now two unopened doors and the prize is behind one of them. Shouldn't there simply be a 0.5 (50%) probability of the prize being behind either of these doors, and therefore it doesn't matter whether you stay with your original choice or switch?

Let's look at all the possible scenarios. Assume that that your first guess is door B. (It doesn't matter which door you guess first, the answer always comes out the same.)

1. If the prize is behind door A, then the host must tell you that the prize is not behind door C.
2. If the prize is behind door B, then the host can tell you either that the prize is not behind door A or that the prize is not behind door C.
3. If the prize is behind door C, then the host must tell you that the prize is not behind door A.

Since each of the above three situations is equally likely, they each have a probability of 1/3.

In situation 1, if you stay with your first choice (door B), you lose. You have the option of switching to door A. If you switch to door A, you win.

In situation 2, if you stay with your first choice (door B), you win. If the host tells you that the prize is not behind door A and you switch to door C, you lose. Also, if the host tells you that the prize is not behind door C and you switch to door A, you lose.

In situation 3, if you stay with your first choice (door B), you lose. You have the option of switching to door C. If you switch to door C, you win.

At this point, Table 1.8 is in order. Remember, your first choice was door B.

TABLE 1.8. Monty Hall Game, Door B Being the First Choice

Prize Location	Remaining Doors	Definite Losers	Stay With Choice B	Switch
A	A, C	C	Lose	Win
B	A, C	A & C	Win	Lose
C	A, C	A	Lose	Win

It appears that if you stay with your first choice, you only win in one of three equally likely situations, and therefore your probability of winning is exactly 1/3. This shouldn't really surprise you. The probability of correctly guessing one door out of three is 1/3, and there's not much more that you can say about it.

On the other hand, if your only options are to stay with your first choice or to switch to the other unopened door, then your probability of winning if you switch must be $1 - 1/3 = 2/3$. There's no getting around this: Either you win or you lose and the probability of winning plus the probability of losing must add up to the certain event—that is, to a probability of 1.

What just happened? What has happened that's different from having just flipped a coin five times, having gotten five heads, and wondering about the sixth flip?

In the coin flipping example, neither the probabilities of the different possibilities or your knowledge of these probabilities changed after five coin flips. In other words, you neither changed the situation nor learned more about the situation. (Obviously, if someone took away the original coin and replaced it with a two headed coin, then expectation values for future flips would change.) In this game-show example, only your knowledge of the probabilities changed; you learned that the probability of the prize being behind one specific door was zero. This is enough, however, to make it possible that the expected results of different actions on your part will also change.

In a later chapter we'll look at another very unintuitive situation: a combination of games, known as "Parrondo's Paradox," where jumping randomly between two losing games creates a winning game because one of the losing games involves looking back at how much you've already won or lost.

THE CHECKERBOARD {DEALING WITH ONLY PART OF THE DATA SET}

Imagine an enormous checkerboard: The board is 2000 squares wide by 2000 squares long. There are $2000 \times 2000 = 4{,}000{,}000$ (four million) squares on the board. Assume that each square has an indentation that can capture a marble.

I'll treat the board as an imaginary map and divide it up into regions, each region containing 1000 indentations. The regions themselves need not be square or even rectangular, so long as each region contains exactly 1000 indentations. There are $4{,}000{,}000/1000 = 4000$ of these regions on the board. There is nothing magic in the choice of any of these numbers. For the purposes of the example, I simply need a large total area (in this case 4 million squares) divided into a lot of small regions (in this case 4000 regions). Also, the regions do not all have to be the same size, it just makes the example easier to present.

Now, I'll lay this checkerboard flat on the ground and then climb up to the roof of a nearby building. The building must be tall enough so that the checkerboard looks like a small dot when I look down. This is important because it assures that if I were to toss a marble off the roof, it would land randomly somewhere on or near the checkerboard, but that I can't control where. I then start tossing marbles off the roof until 40,000 marbles have landed on the checkerboard and are trapped in 40,000 indentations. This is, admittedly, a very impractical experiment. I won't worry about that, however, because I'm not really planning to perform the experiment. I just want to describe a way of picturing the random scattering of 40,000 objects into 4,000,000 possible locations. The choice of 40,000 objects isn't even a critical choice, it was just important to choose a number that is a small fraction of 4,000,000 but is still a fairly large number of objects. In any case, when I am done we see that the fraction of the number of indentations that I have filled is exactly

$$\frac{40{,}000}{4{,}000{,}000} = \frac{1}{100} = 0.01 = 1\%$$

Now let's take a close look at a few of the 4000 regions, each of which has 1000 indentations. Since 1% of the indentations are filled with marbles, we would expect to see 1% of 1000, or

$$0.01 \times 1000 = 10$$

marbles in each region. On the average, over the 4000 regions, this is exactly what we must see—otherwise the total number of marbles would not be 40,000. However, when we start looking closely, we see something very interesting[5]: Only about 500 of the regions have 10 marbles.[6] About 200 of the regions have 14 marbles, and about 7 of the regions have 20 marbles. Also, about 9 of the regions have only 2 marbles. Table 1.9 tabulates these results.

What Table 1.9 is showing us is that while the most likely situation, in this case 10 marbles per region, will happen more often than any other situation, the most likely situation is not the only thing that will happen (just like in the coin flip game). The results are distributed over many different situations, with less likely situations happening less often. In order to predict what we see from this experiment, therefore, we not only need to know the most likely result, but also need to know something about how a group of results will be distributed among all possible results. These *probability distributions* will be a subject of the next chapter.

[5] At this point I am not attempting to explain how the observations of "what we actually see" come about. This will be the subject of the chapter on binomial distributions (Chapter 6).

[6] I say "about" because it is very unlikely that these numbers will repeat exactly if I were to clear the board and repeat the experiment. How the results should be expected to vary over repeated experiments will also be the subject of a later chapter.

TABLE 1.9. Expected Number of Marbles in Regions on Giant Checkerboard

Number of Regions	Number of Marbles	Number of Regions	Number of Marbles
2	1	292	13
9	2	208	14
30	3	138	15
74	4	86	16
150	5	50	17
251	6	28	18
360	7	14	19
451	8	7	20
502	9	3	21
503	10	2	22
457	11	1	23
381	12		

Before leaving this example, let's play with the numbers a little bit and see what we might learn. Since the most likely result (10 marbles per region) occurs only about 500 times out of 4000 opportunities, some other results must be occurring about 3500 times out of these 4000 opportunities. Again, we have to be very careful what we mean by the term "most likely result." We mean the result that will probably occur *more times than any other result* when we look at the whole checkerboard. The probability that the most likely result will not occur in any given region is about

$$\frac{3500}{4000}=\frac{7}{8}=0.875=87.5\%$$

Putting this in gambling terms, there are 7 to 1 odds against the most likely result occurring in a given region.

Now, suppose someone is interested in the regions that have at least 20 marbles. From the table we see that there are 13 of these regions. It wouldn't be surprising if a few of them are near an edge of the board. Let's imagine that this person locates these regions and takes a picture of each of them. If these pictures were shown to you and you are not able to look over the rest of the board yourself, you might tend to believe the argument that since there are some regions near the edge of the board that have at least twice the average number of marbles, then there must be something about being near the edge of the board that "attracts marbles." There are, of course, many regions that have less than 1/2 the average number of marbles, and some of these are probably near the edge of the board too, but this information is rarely mentioned (it just doesn't make for good headlines). Instead, we see "Cancer Cluster Near Power Lines" and similar statements in the newspapers.

It's hard to generalize as to whether the reporters who wrote the story intentionally ignored some data, unintentionally overlooked some data, didn't understand what they were doing by not looking at all the data, or were just looking to write a good story at the cost of total, complete truthfulness.

In all fairness, I am now leaving the realm of mathematics and meddling in the realms of ecology, public health, and so on. There are indeed unfortunate histories of true disease clusters near waste dumps, and so on. The point that must be made, over and over again, however, is that you cannot correctly spot a pattern by "cherry picking" subsets of a large data set. In the case of power lines near towns, when you look at the entire checkerboard (i.e., the entire country), you find high disease clusters offset by low disease clusters and when you add everything up, you get the average. If there were high disease clusters which were truly nonrandom, then these clusters would not be offset by low disease clusters, and you would get a higher-than-average disease rate when you add everything up. Also, you must be prepared to predict the variability in what you see; for example, a slightly higher-than-average total might just be a random fluctuation, sort of like flipping a coin and getting 5 heads in a row. The World Health Organization maintains a database on their website in the section on Electromagnetic Fields, where you can get a balanced perspective of the power-line and related studies.

This same issue shows up over and over again in our daily lives. We are bombarded with everything from health food store claims to astrological predictions. We are never shown the results of a large study (i.e., the entire checkerboard). The difficult part here is of course that some claim might be absolutely correct. The point, however, is that we are not being shown the information necessary to see the entire picture, so we have no way of correctly concluding whether or not a claim is correct based upon looking at the entire picture or just anecdotal without further investigation. And, of course in the area of better health or longer life or . . . , we are genetically programmed to bet on the pattern we are shown, just in case there are man-eating tigers lurking in the brush.

It is not uncommon to be confronted with a situation where it is either impossible or very highly impractical to study the entire data set. For example, if we are light bulb manufacturers and we want to learn how long our light bulbs last, we could ask every customer to track the lifetime of every bulb and report back to us, but we know that this is not going to happen. We could run all of our light bulbs ourselves until they burn out, but this is not a very good business plan. Instead, we take a representative sample of the bulbs we manufacture, run them ourselves until they burn out, and then study our data. In a later chapter I will discuss what it takes for a sample to be representative of the entire population. We need to know what fraction of the manufactured bulbs we must choose as samples, what rules we need to follow in selecting these samples, and just how well the results of studying the sample population predict the results for the entire population.

I hope that these examples and the explanations have created enough interest that you'd like to continue reading. The next two chapters will present most of the basic mathematics and definitions needed to continue. Then starting in Chapter 4 I will concentrate on examples of how random variables and probability affect many of the things we do and also how they are at the basis of many of the characteristics of the world we live in.