PROBABLY NOT

Future Prediction Using Probability and Statistical Inference

Lawrence N. Dworsky
Phoenix, AZ

A JOHN WILEY & SONS, INC. PUBLICATION

PROBABLY NOT

PROBABLY NOT

Future Prediction Using Probability and Statistical Inference

Lawrence N. Dworsky
Phoenix, AZ

A JOHN WILEY & SONS, INC. PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Dworsky, Lawrence N., 1943–
Probably not : future prediction using probability and statistical inference / Lawrence N. Dworsky.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-18401-1 (pbk.)
1. Prediction theory. 2. Probabilities. 3. Mathematical statistics. I. Title.
QA279.2.D96 2008
519.5′4—dc22

2007050155

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To my grandchildren, Claudia and Logan, and to grandchildren everywhere—probably the best invention of all time.

CONTENTS

PREFACE

For as long as I can remember, I have been interested in how well we know what we say we know, how we acquire data about things, and how we react to and use this data. I was surprised when I first realized that many people are not only not interested in these things, but are actually averse to learning about them. It wasn't until fairly recently that I concluded that we seem to be genetically programmed, on one hand, to intelligently learn how to acquire data and use it to our advantage but also, on the other hand, to stubbornly refuse to believe what some simple calculations and/or observations tell us.

My first conclusion is supported by our march through history, learning about agriculture and all the various forms of engineering and using this knowledge to make life better and easier. My latter conclusion comes from seeing all the people sitting on stools in front of slot machines at gambling casinos, many of whom are there that day because the astrology page in the newspaper told them that this was "their day."

This is a book about probability and statistics. It's mostly about probability, with just one chapter dedicated to an introduction to the vast field of statistical inference.

There are many excellent books on this topic available today. I find that these books fall into two general categories. One category is textbooks. Textbooks are heavily mathematical with derivations, proofs and problem sets, and an agenda to get you through a term's course work. This is just what you need if you are taking a course.

The other category is books that are meant for a more casual audience—an audience that's interested in the topic but isn't interested enough to take a course. We're told today that people have "mathephobia," and the books that

appeal to these people try very hard to talk around the mathematics without actually presenting any of it. Probability and statistics are mathematical topics. A book on these subjects without math is sort of like a book on French grammar without any French words in it. It's not impossible, but it sure is doing things the hard way.

This book tries to split the difference. It's not a textbook. There is, however, some math involved. How much? Some vague reminiscences about introductory high school algebra along with a little patience in learning some new notation should comfortably get you through it. You should know what a fraction is and recognize what I'm doing when I add or multiply fractions or calculate decimal equivalents. Even if you don't remember how to do it yourself, just realizing what I'm doing and accepting that I'm probably doing it right should be enough. You should recognize a square root sign and sort of remember what it means. You don't need to know how to calculate a square root—these days everybody does it on a pocket calculator or a computer spreadsheet anyway. You should be able to read a graph. I review this just in case, but a little prior experience helps a lot. In a few cases some elementary calculus was needed to get from point A to point B. In these cases I try to get us all to point A slowly and clearly and then just say that I needed a magic wand to jump from A to B and that you'll have to trust me.

If you thumb through the book, you'll see a few "fancy" formulas. These are either simply shorthand notations for things like repeated additions, which I discuss in great detail to get you comfortable with them, or in a few cases some formulas that I'm quoting just for completeness but that you don't need to understand if you don't want to.

As I discuss in the first chapter, probability is all about patterns of things such as what happens when I roll a pair of dice a thousand times, or what the life expectancies of the population of the United States looks like, or how a string of traffic lights slows you down in traffic. Just as a course in music with some discussions of rhythm and harmony helps you to "feel" the beauty of the music, a little insight into the mathematics of the patterns of things in our life can help you to feel the beauty of these patterns as well as to plan things that are specifically unpredictable (when will the next bus come along and how long will I have to stand in the rain to meet it?) as best possible.

Most popular science and math books include a lot of biographical information about the people who developed these particular fields. This can often be interesting reading, though quite honestly I'm not sure that knowing how Einstein treated his first wife helps me to understand special relativity.

I have decided not to include biographical information. I often quote a name associated with a particular topic (Gaussian curves, Simpson's Paradox, Poisson distribution) because that's how it's known.

Probabilistic considerations show up in several areas of our lives. Some we get explicitly from nature, such as daily rainfall or distances to the stars. Some we get from human activities, including everything from gambling games to manufacturing tolerances. Some come from nature, but we don't see

them until we "look behind the green curtain." This includes properties of gases (e.g., the air around us) and the basic atomic and subatomic nature of matter.

Mathematical analyses wave the banner of *truth*. In a sense, this is deserved. If you do the arithmetic correctly, and the algorithms, or formulas, used are correct, your result is *the* correct answer and that's the end of the story. Consequently, when we are presented with a conclusion to a study that includes a mathematical analysis, we tend to treat the conclusion as if it were the result of summing a column of numbers. We believe it.

Let me present a situation, however, where the mathematics is absolutely correct and the conclusion is absolutely incorrect. The mathematics is simple arithmetic, no more than addition and division. There's no *fancy stuff* such as probability or statistics to obfuscate the thought process or the calculations. I've changed the numbers around a bit for the sake of my example, but the situation is based upon an actual University of California at Berkeley lawsuit.

We have a large organization that is adding two groups, each having 100 people, such as two new programs at a school. I'll call these new programs A and B.

Program A is an attractive program and for some reason is more appealing to women than it is to men. 600 women apply; only 400 men apply. If all the applicants were equally qualified, we would expect to see about 60 women and 40 men accepted to the 100 openings for the program. The women applicants to this program tend to be better qualified than the men applicants, so we end up seeing 75 women and 25 men accepted into program A. If you didn't examine the applications yourself, you might believe that the admissions director was (unfairly) favoring women over men.

Program B is not as attractive a program and only 100 people apply. It is much more attractive to men than it is to women: 75 men and 25 women apply. Since there are 100 openings, they all get accepted.

Some time later, there is an audit of the school's admission policies to see if there is any evidence of unfair practices, be they sexual, racial, ethnic, whatever. Since the new programs were handled together by one admissions director, the auditor looks at the books for the two new programs as a group and sees that:

600 + 25 = 625 women applied to the new programs. 75 + 25 = 100 women were accepted. In other words, 100/625 = 16% of the women applicants were accepted to the new programs.

400 + 75 = 475 men applied to the new programs. 25 + 75 = 100 men were accepted. In other words, 100/475 = 21% of the men applicants were accepted to the new programs.

The auditor then reviews the qualifications of the applicants and sees that the women applicants were in no way inferior to the men applicants; in fact it's the opposite. The only plausible conclusion is that the programs' admissions director favors men over women.

The arithmetic above is straightforward and cannot be questioned. The flaw lies in how details get lost in summarization—in this case, looking only at the totals for the two programs rather than keeping the data separate. I'll show (in Chapter 13) how a probabilistic interpretation of these data can help to calculate a summary correctly.

My point here, having taken the unusual step of actually putting subject matter into a book's preface, is that mathematics is a tool and only a tool. For the conclusion to be correct, the mathematics along the way must be correct, but the converse is not necessarily true.

Probability and statistics deals a lot with examining sets of data and drawing a conclusion—for example, "the average daily temperature in Coyoteville is 75 degrees Fahrenheit." This sounds like a great place to live until you learn that the temperature during the day peaks at 115 degrees while at night it drops to 35 degrees. In some cases we will be adding insight by summarizing a data set, but in some cases we will be losing insight.

My brother-in-law Jonathan sent me the following quote, attributing it to his father. He said that I could use it if I acknowledge my source: Thanks, Jonathan.

"The average of an elephant and a mouse is a cow, but you won't learn much about either elephants or mice by studying cows." I'm not sure exactly what the arithmetic in this calculation would look like, but I think it's a memorable way of making a very good point.

I could write a long treatise on how bad conclusions have been reached because the people who had to draw the conclusions just weren't looking at all the data. Two examples that come to mind are (1) the Dow silicone breast implant lawsuit where a company was put out of business because the plaintiffs "demonstrated" that the data showed a link between the implants and certain serious disease and (2) the crash of the space shuttle Challenger where existing data that the rubber O-rings sealing the liquid hydrogen tanks get brittle below a certain temperature somehow never made it to the table.

The field of probability and statistics has a very bad reputation ("Lies, Damned Lies, and Statistics"[1]). It is so easy to manipulate conclusions by simply omitting some of the data, or to perform the wrong calculations correctly, or to misstate the results—any and all of these possibly innocently—because some problems are very complicated and subtle. I hope the materials to follow show what information is needed to draw a conclusion and what conclusion(*s*) can and can't be drawn from certain information. Also I'll show how to reasonably expect that sometimes, sometimes even inevitably, as the bumper stickers say, *stuff* happens.

[1] This quote is usually attributed to Benjamin Disraeli, but there seems to be some uncertainty here. I guess that, considering the book you're now holding, I should say that "There is a high probability that this quote should be attributed to Benjamin Disraeli."

I spend a lot of time on simple gambling games because, even if you're not a gambler, there's a lot to be learned from the simplest of random events—for example, the result of coin flips.

I've also tried to choose many examples that you don't usually see in probability books. I look at traffic lights, waiting for a bus, life insurance, scheduling appointments, and so on. What I hope to convey is that we live in a world where so many of our daily activities involve random processes and the statistics involved with them.

Finally, I introduce some topics that show how much of our physical world is based on the consequences of random processes. These topics include gas pressure, heat engines, and radioactive decay. These topics are pretty far from things you might actually do such as meeting a friend for lunch or counting birds in the woods. I hope you'll find that reading about them will be interesting.

One last comment: There are dozens and dozens of clever probability problems that make their way around. I've included several of these (the shared birthday, the prize behind one of three doors, etc.) where appropriate and I discuss how to solve them. When first confronted with one of these problems, I inevitably get it wrong. In my own defense, when I get a chance to sit down and work things out carefully, I (usually) get it right. This is a tricky subject. Maybe that's why I find it to be so much fun.

ACKNOWLEDGMENTS

My wife, Suzanna, patiently encouraged me to write this book for the many months it took me to actually get started. She listened to my ideas, read and commented on drafts and perhaps most importantly, suggested the title.

My daughter Gillian found the time to read through many chapter drafts and to comment on them. My son-in-law Aaron and I had many interesting conversations about the nature of randomness, random sequences of numbers, and the like that clarified my thoughts significantly.

My friends Mel Slater and Arthur Block and my brother-in-law Richie Lowe also found the time to read and comment on chapter drafts.

The folks at John Wiley were very encouraging and helpful from the start.

I am grateful to all of you.

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.