# Probabilities: The Little Numbers That Rule Our Lives

# Probabilities: The Little Numbers That Rule Our Lives

Peter Olofsson

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

# Preface

This book is about those little numbers that we just cannot escape. Try to remember the last day you didn't hear at least something about probabilities, chance, odds, randomness, risk, or uncertainty. I bet it's been a while. In this book, I will tell you about the mathematics of such things and how it can be used to better understand the world around you. It is not a textbook though. It does not have little colored boxes with definition or theorems, nor does it contain sections with exercises for you to solve. My main purpose is to entertain you, but it is inevitable that you will also learn a thing or two. There are even a few exercises for you, but they are so subtly presented that you might not even notice until you have actually solved them.

The spousal thanks is always more than a formality. I thank Αλκμήνη for putting up with irregular work hours and everything else that comes with writing a book, but also for help with Greek words and for reminding me of some of my old travel stories that you will find in the book. I am deeply grateful to Professor Olle Häggström at Chalmers University of Technology in Göteborg, Sweden. He has read the entire manuscript, and his comments are always insightful, accurate, and clinically free from unnecessary politeness. If you find something in this book that strikes you as particularly silly, chances are that Mr. Häggström has already pointed it out to me but that I decided to keep it for spite. I have also received helpful comments from John Haigh at the

University of Sussex, Steve Quigley at Wiley, and from an anonymous referee. Thanks also to Kris Parrish and Susanne Steitz at Wiley, to Sheree Van Vreede at Sheree Van Vreede Publications Services for excellent copyediting, and to Amy Hendrickson at Texnology Inc. for promptly and patiently answering my LaTeX questions.

A large portion of this book was written during the tumultuous Fall of 2005. Our move from Houston to New Orleans in early August turned out to be a masterpiece of bad timing as Hurricane Katrina hit three weeks later. We evacuated to Houston, and when Katrina's sister Rita approached, we took refuge in the deserts of West Texas and New Mexico. Sandstorms are so much more pleasant than hurricanes! However, it was also nice to return to New Orleans in January 2006; the city is still beautiful, and its chargrilled oysters are unsurpassed. I am grateful to many people who housed us and helped us in various ways during the Fall and by doing so had direct or indirect impact on this book. Special thanks to Kathy Ensor & Co. at the Department of Statistics at Rice University in Houston and to Tom English & Co. at the College of the Mainland in Texas City for providing me with office space. Finally, thanks to Professor Peter Jagers at Chalmers University of Technology, who as my Ph.D. thesis advisor once in a distant past wisely guided me through my first serious encounters with probabilities, those little numbers that rule our lives.

PETER OLOFSSON
www.peterolofsson.com

*San Antonio, 2014*

# Contents

1

# Computing Probabilities: Right Ways and Wrong Ways

2

# Surprising Probabilities: When Intuition Struggles

# 3

# Tiny Probabilities: Why Are They So Hard to Escape?

# 4

# Backward Probabilities: The Reverend Bayes to Our Rescue

# 5

# Beyond Probabilities: What to Expect

## GREAT EXPECTATIONS

In the previous chapters, I have several times talked about what happens "on average" or what you can "expect" in situations where there is randomness involved. For example, on page **??**, it was pointed out that the parameter $\lambda$ in the Poisson distribution is the *average* number of occurrences. I have mentioned that each roulette number shows up on *average* once every 38 times and that you can *expect* two sixes if you roll a die 12 times. The time has come to make this discussion exact, to look beyond probabilities and introduce what probabilists call the *expected value*. This single number summarizes an experiment, and in order to compute an expected value, you need to know all possible outcomes and their respective probabilities. You then multiply each value by its probability and add everything up. Let us do a simple example.

Roll a fair die. The possible outcomes are the numbers 1 through 6, each occurring with probability $1/6$, and by what I just described, we get

$$1 \times 1/6 + 2 \times 1/6 + 3 \times 1/6 + 4 \times 1/6 + 5 \times 1/6 + 6 \times 1/6 = 3.5$$

as the expected value of a die roll. You may notice that the term "expected" is a bit misleading because you certainly do not expect to get 3.5 when you roll the die. Think instead of the expected value as the expected *average* in a large number of rolls of the die. For example, if you roll the die five times and get the numbers 2, 3, 1, 5, 3, the average is $(2+3+1+5+3)/5 = 2.8$. If you roll another five times and get 2, 5, 1, 4, 5, the average over the ten rolls is $31/10 = 3.1$. As you keep going, rolling over and over and computing consecutive averages, you can expect these to settle in toward 3.5. I will elaborate more on this interpretation and make it precise in the next chapter. You can also think of the "perfect experiment" in which the die is rolled six times and each side shows up exactly once. The average of the six outcomes in the perfect experiment is 3.5, and this is the expected value of a die roll.

In the casino game *craps*, two dice are rolled and their sum recorded. What is the expected value of this sum? The 11 possible values are 2, 3,..., 12, but these are not all equally likely so we must figure out their probabilities. In order to get 2, both dice must show 1 and the probability of this is $1/36$. In order to get 3, one die must show 1 and the other 2 and as there are two dice that can each play the role of "one" or "the other," there are two outcomes that give the sum 3. The probability to get 3 is therefore $2/36$. To get 4, any of the three combinations 1–3, 2–2, or 3–1 will do, so the probability is $3/36$, and so on and so forth. The outcome 7 has the highest probability, $6/36$, and from there the probabilities start to decline down to $1/36$ for the outcome 12 (consult Figure **??** on page **??** if you feel uncertain about these calculations). Now add the outcomes multiplied by their probabilities to get

$$2 \times 1/36 + 3 \times 3/36 + \cdots + 12 \times 1/36 = 7$$

as the expected sum of two dice. This time, the expected value is a number than you *can* get as opposed to the 3.5 with one die. Unfortunately, it still does not mean that you actually expect to get 7 in each roll or you would expect to leave the casino a wealthy person, 7 being a winning number in craps. It only refers to what you can expect on average in the long run.

Note that the expected value of the sum of two dice, 7, equals twice the expected value of the outcome of one die, 3.5. This is no coincidence. Expected values have the nice property of being what is called *additive*, which means that we did not have to do the calculation we did above for the two dice. Instead, we could just have said that as we roll two dice and each has the expected value 3.5, the expected value of the sum is $3.5 + 3.5 = 7$. This is convenient. If you roll 100 dice, you know that the expected sum is 350 without having to figure out how to combine the outcomes of 100 dice to get the sum 298 or 583 (but feel free to try it, at least it will keep you out of trouble).

Expected values are more than additive; they are also *linear*, which is a more general concept. In addition to additivity, linearity means that if you multiply each outcome by some constant, the expected value is multiplied by the same constant. For example, roll a die and double the outcome. The expected value of the doubled outcome is then twice the expected value of the outcome of the roll, $2 \times 3.5 = 7$. Note that the expected values are the same when you *double* the outcome of *one* roll and when you *add* the outcomes of *two* rolls. The actual experiments are different though. In the first case, the possible values are the 6 even numbers 2, 4,..., 12; in the second, the 11 numbers 2, 3, ..., 12.

To illustrate the convenience of linearity, suppose that you construct a random rectangle by rolling three dice. The first determines one side, and the sum of the other two determines the other side. What is the expected circumference of the random rectangle? There are 216 different outcomes of the three dice. The smallest possible rectangle measures one by two, has circumference 6, and probability $1/216$ because there is only one way to get it: $(1,1,1)$. The largest rectangle measures 6 by 12, has circumference 36, and likewise probability $1/216$. In between these, there is a range of possibilities with different probabilities. For example, you can get circumference 8 in three different ways: $(1,1,2)$, $(1,2,1)$, and $(2,1,1)$, so circumference 8 has probability $3/216$ (these three rectangles have dimensions $1\times3$, $1\times3$, and $2\times2$, respectively). To compute the expected circumference, however, you do not need to figure out all these outcomes and their probabilities. Simply note that the circumference is twice one side plus twice the other; that the sides have expected lengths 3.5 and 7, respectively; and apply linearity to get the expected circumference $2 \times 3.5 + 2 \times 7 = 21$. Linearity is a very convenient property indeed.

You may have noticed that the expected values in our examples thus far have been right in the middle of the range of possible outcomes. The midpoint of the numbers 1, 2,..., 6 is 3.5; the midpoint of 2, 3,..., 12 is 7; the midpoint of 100, 101,..., 600 is 350; and the midpoint of the rectangle circumference values is $(6 + 36)/2 = 21$. All these examples have in common that the probability distributions are *symmetric*. If you roll one die, start from 3.5, which is in the middle and step outward in both directions: 3 and 4 have the same probabilities; 2 and 5 have the same probabilities; and 1 and 6 have the same probabilities. Of course, in this particular case, *all* outcomes have the same probabilities, $1/6$, so it may be more interesting to look at the sum of two dice instead. Here, 7 is in the middle and has probability $6/36$. One step out we find 6 and 8, which both have probability $5/36$. Continue like this until you hit the last two outcomes 2 and 12, each with probability $1/36$. So in these cases, you could actually have found the expected value simply by computing the average of the possible outcomes: The average of 1, 2,..., 6 is 3.5, and the average of 2, 3,..., 12 is 7. This is not always the case though, and here is another dice example to prove it. Roll two dice and record the *largest* number. What is the expected value?

The largest number can be anything from 1 to 6, but these are not equally likely, nor are the probabilities distributed symmetrically, and it is probably clear that the largest value is expected to be more than 3.5. To find the expected value, we need to first compute the probabilities. The only case in which the largest number equals 1 is when both dice show 1, and this has probability $1/36$. Three cases give the largest number 2: 1–2, 2–1, and 2–2, and the probability is thus $3/36$. Continue like this until you reach 6, which is the largest number in 11 cases and has probability $11/36$ (you might again find it helpful to consider Figure **??**). If you are into math formulas, the probability that the largest number equals $k$ is $(2 \times k - 1)/36$ for $k$ ranging from 1 to 6. At any rate, the expected value is now computed as

$$1 \times 1/36 + 2 \times 3/36 + \cdots + 6 \times 11/36 \approx 4.5$$

rounded to one decimal. I leave it up to you to demonstrate that the expected *smallest* number is $\approx 2.5$ (obvious without calculations?).

Here is another example of asymmetric probabilities, which also involves negative numbers in a natural way. You play roulette and bet \$1 on the number 29. What is your expected gain? There are two possibilities: With probability $1/38$, number 29 comes up and you win \$35, and with probability $37/38$, some other number comes up and you lose your dollar. If we agree to describe a loss as a negative gain, your gain can therefore be either 35 or $-1$. There is no problem with having a negative number, and the expected value of your gain is computed just like before:

$$35 \times 1/38 + (-1) \times 37/38 = -2/38 \approx -0.0526$$

an expected loss of about 5 cents. Again we have a case in which the expected value cannot actually occur but must be interpreted as a long-term average. In the long run, each number comes up once every 38 spins, so assume that this is *exactly* what happens; the numbers come up perfectly in order: 00, 0, 1, 2, ..., 38 and you bet \$1 on 29 each time, wagering a total of \$38. You will then lose \$37 and win \$35 (and keep the dollar you bet on 29), a total loss of \$2 out of the \$38.

You often see people at the roulette tables betting on several different numbers, sometimes covering almost the entire table. Although this certainly increases your chances of winning in a single spin, it does nothing to improve your expected long-term losses. Indeed, you lose 5 cents per dollar on each single number, so if you for example bet \$1 on each of ten different numbers, additivity of expected values tells you that you can expect to lose on average 50 cents. Regardless of betting strategy, the casino takes on average 5 cents out of every dollar you risk, which does not sound like much but is enough to give them a very good profit.

As an exercise, let us compute the expected gain in the more innocent chuck-a-luck. In this game, you wager \$1, three dice are rolled, and your win depends on the number of 6s. If there is one 6, you win \$1; if there are two 6s, you win \$2; and if there are three 6s, you win \$3. Only if there are no 6s do you lose your \$1. On page **??**, we saw that the probability that you win something is 0.42. Thus, you lose your \$1 with probability 0.58, but if you win, you may win more than \$1 so it is not immediately obvious that the game is stacked against you. The probabilities to get zero, one, two, and three 6s are

$$\begin{array}{lllll} P(\text{no 6s}) & = (5/6)^3 & = 125/216 & \approx & 0.58 \\ P(\text{one 6}) & = 3 \times 1/6 \times (5/6)^2 & = 75/216 & \approx & 0.35 \\ P(\text{two 6s}) & = 3 \times (1/6)^2 \times 5/6 & = 15/216 & \approx & 0.07 \\ P(\text{three 6s}) & = (1/6)^3 & = 1/216 & \approx & 0.005 \end{array}$$

where the “3” in the two middle probabilities is there because the one die that shows different from the other can be any of the three (in fact, the number of 6s has a binomial distribution, which we discussed in Chapter 1). The decimal numbers above do not add up to 1 as they should, but that is only because they are rounded. Let us now compute the expected gain. Your gain equals the number of 6s if you get any, and otherwise, it is $-1$. The expected gain in chuck-a-luck is

$$(-1) \times 125/216 + 1 \times 75/216 + 2 \times 15/216 + 3 \times 1/216 \approx -0.08$$

that is, an expected loss of about 8 cents per \$1 wagered. From a financial point of view, you’re worse off than at the roulette table. Again you can think of what would happen in the ideal run where each of the 216 possible outcomes of the three dice comes up exactly once. You then win \$3 once, \$2

15 times, \$1 75 times, and lose \$1 125 times, for a total loss of \$17 of your \$216 wagered.

Suppose that you try another kind of gambling: stock investments. A friend tells you that a particular mutual fund is equally likely to either go up 50% or down 40% each year for the next few years to come. If you invest \$1,000, how much can you expect to have after two years?

First consider one year. After the first year you are equally likely to have \$1,500 or \$600, and the average of these is \$1,050. In general, the average of a 50% gain and a 40% loss is a 5% gain, so you can expect to gain 5% each year. After two years, your expected fortune is therefore \$1,000 $\times$ 1.05 $\times$ 1.05 $=$ \$1,102.50. On the other hand, as your fortune is equally likely to increase as it is to decrease each year and there are two years, you can expect it to go down one year and up the other. Regardless of which of these years that comes first, your fortune will be \$1,000 $\times$ 1.50 $\times$ 0.60 $=$ \$900. This seems conflicting. How can you expect your fortune both to increase and to decrease?

It depends on what you mean by "expect." The expected value of your fortune after two years is certainly \$1,102.50. There are four equally likely scenarios for the two years: up–up, up–down, down–up, and down–down, leading to fortunes of \$2,250, \$900, \$900, and \$360, respectively, and the average of these is \$1,102.50. However, if you instead compute the expected number of "good years," this number is one and the *most likely* scenario is one good and one bad year, which makes \$900 the most likely value of your fortune. The most likely value, in this case \$900, is called the *mode* or *modal value*. It is up to you which of the two measures of your fortune you think makes most sense. Note that although your *expected* fortune increases, the *actual* fortune only increases if there are two good years of which there is a 25% chance. If you compare this investment scheme with one that gives a fixed 5% interest each year, the two are on average equally good and equally likely to be ahead after a year. However, the fixed interest scheme has a 75% chance of being ahead of the mutual fund after two years. If they compete, it is a fair game year by year but not over two years, somewhat paradoxically. And as the years keep passing by, your expected fortune increases by 5% each year, but under the most likely scenario your fortune instead decreases by 10% every two years. After 20 years, your initial investment of \$1,000 has grown to \$2,653 as measured by expected returns and fallen to \$349 under the most likely scenario. In order for your actual fortune to increase after 20 years, you need at least 12 good years, which has a probability of about 25%.

The rates in the example may not be very realistic but serve as a drastic illustration to the general principle that a decrease is more severe than an increase. For example, if a 50% gain is followed by a 50% loss (or vice versa), this leaves you with a net loss of 25%. The combination of a 10% gain and a 10% loss results in a net loss of 1%, and so on. If equally sized annual gains and losses are equally likely, your expected fortune remains unchanged, but in order for the actual fortune not to decrease, you need more good years than bad. This is still true even if the annual gains tend to be slightly larger than the annual losses (as in the extreme example above).

When it comes to risking money in order to make money, you must of course weigh risk against benefit and considering only the expected gain is not sufficient. You may buy a lottery ticket for the slim chance to win big even though you face an expected loss. But if I offer you the chance to bet $1,000 on a coin toss and pay you $1,100 if you get heads, you might not want to play even with the expected gain of $50. In the long run you would certainly ruin me, but for a single bet you might not be willing to risk $1,000 for the chance of winning that extra $100. You face similar concerns when it comes to investing your money. Should you take a risk on highly uncertain but potentially very profitable stocks, or should you go with the lower risks of mutual funds or bonds? The expected return should play a role in your decision but should definitely not be the sole criterion. Let us look closer at the mathematics of stock markets.

## OPTIONS...

EFFICIENT MARKET HYPOTHESIS

OPTIONS, EUROPEAN, AMERICAN

You have $100 to invest and have three choices: (a) a risk-free bond, (b) a stock currently trading at $100, or (c) an option to buy the same stock. The main question is how much the option should cost and the main assumption is that your expected gain is the same regardless of what you choose to do, assuming this is what the efficient market hypothesis implies. Let us first consider (a), the risk-free bond, and assume the annual interest rate is 5%. Thus, if you buy the bond, you will have $105 a year later. Next, consider (b), the stock that is currently trading at $100. To simplify things considerably, we will assume that the stock price can only take on 2 different values after a year: \$150 or \$50. Although completely unrealistic, the assumption is made to simplify the calculations and illustrate an idea; we will later consider more realistic scenarios. It wold be tempting to assume the stock equally as likely to increase as to decrease, thus making its expected value after a year the same as its current price, \$100, but remember that we have assumed that we will do as well with the stock as with the bond, on average. Thus, the expected value of the stock after a year is \$105 and if we let $p$ denote the probability of an increase, we get the equation

$$100 \times p + 50 \times (1 - p) = 150$$

which we can solve for $p$ to get $p = 0.55$. Thus, there is a 55% chance that the stock increases in value to \$150 and a 45% chance that it decreases to \$50. With these probabilities, you do as well buying the bond as the stock.

What, then, about (c), the option? How much should you pay for it in order to have the same expected fortune of \$105 a year later? If you pay \$ $D$ for the option, after a year you will have a fortune of \$ $(150 - D)$ if the stock has gone up so that you exercise the option to buy the stock. There is a 55% chance of this scenario. If the stock goes down, of which there is a 45% chance, you do not exercise the option and your fortune is \$ $(100 - D)$. We now get the equation

$$(150 - D) \times 0.55 + (100 - D) \times 0.45 = 105$$

which has the solution $D = 22.5$. Thus, the fair value of the option is \$22.50. If somebody offers the option at a lower price, you should buy it rather than the stock itself for a higher expected yield.

So what about this unrealistic assumption that the stock can only take on two possible values after a year? It's all about time scales. Note that a change from \$100 to \$150 is a 50% increase, and a change from \$100 to \$50 is a 50% decrease. Now, rather than a year, consider changes over two 6-month periods such that we can still get a 50% increase or decrease after a year. If we want to keep percentual changes constant over the two 6-month periods, it turns out that we must let increases be by about 22.5% and decreases by about 29.3%. Thus, after the first 6-month period, the stock is worth either \$122.50 (22.5% increase) or \$70.70 (29.3% decrease). In the first case, if there is another 22.5% increase, the stock is worth \$150 (rounded to the nearest dollar), a 50% increase over the year. In the case of two consecutive drops in price, it is worth 70.7% of \$70.70 which is \$50 (again rounded), a 50% decrease over the year.

Note that we now also have the possibility of an increase followed by a decrease, or vice versa, over the two 6-month periods. In each of these cases, the stock will be worth about $87 after a year ($100 \times 1.225 \times 0.707 = 86.61$). Thus, we now have 3 possible values after a year rather than 2, and by choosing probabilities carefully, we can makje the expected value equal to the $105 we would get with the risk-free bond. Now, 3 is incrementally better than 2 but still, of course, highly unrealistic. But you understand where this is going. Instead of 6-month periods, consider months. Or weeks. Or days, hours, minutes, seconds, milliseconds, and so on until you can imagine (mathematicians are good at this) that the stock price moves continuously in time, in such a way that the expected value after a year is $105. Here we get into the highly advanced territory of *stochastic differential equations* which is the necessary environment for the famous (some would say infamous) *Black-Scholes formula* for option pricing.

**Something about transaction costs?**

## BLOOD AND THEORY

END OF RED TEXT

Careful consideration of expected values can save time and money as the next example illustrates. During World War II, millions of American draftees had their blood drawn to be tested for syphilis, a disease that was expected to be detected in a few thousand individuals. Analyzing the blood samples was a time-consuming and expensive procedure, and a Harvard economist, Robert Dorfman, came up with a clever idea. Instead of testing each individual, he suggested, divide the draftees into groups, draw their blood, and mix some blood from everybody in the group to form a *pooled* blood sample. If the pooled sample tests negative, the whole group is declared healthy, and if it tests positive, each individual sample is tested separately. The point is of course that entire groups can be declared healthy by just one blood sample analysis. The same idea can be used for any disease that is rare and where large populations need to be screened. Let us look at the mathematics of pooled blood samples.

Denote the size of the group by $n$ and the probability that an individual has the disease by $p$.[1] Additional tests must be done if *somebody* has the disease, and because *somebody* is the opposite of *nobody*, this is a case for Trick Number One. The probability that an individual does not have the disease is $1 - p$, and assuming independence between individuals, the probability that nobody has the disease is $(1 - p)^n$. Finally, the probability that somebody has the disease is $1 - (1 - p)^n$, and this is then the probability that the pooled sample tests positive; in which case, $n$ additional individual tests are done. After the first test, with probability $(1 - p)^n$, there are no additional tests, and with probability $1 - (1 - p)^n$, there are $n$ additional tests. The expected number of tests with the pooling method is therefore

$$1 + n \times (1 - (1 - p)^n)$$

where the first 1 is there because one test must always be done and the term $0 \times (1 - p)^n$ that should formally be added was ignored because it equals 0. Now compare this expected value with the $n$ tests that are done if all samples are tested individually. Let us put in some values, for example, $n = 20$ and $p = 0.01$. Then $1 - p = 0.99$, and the expected number of tests is

$$1 + 20 \times (1 - 0.99^{20}) \approx 4.6$$

which is certainly preferred over the 20 tests that would have to be done individually. Note also that even if the pooled blood sample is positive, very little is lost because the pooling method then requires a total of 21 tests instead of 20, only one test more (and there is no need to draw more blood, what was drawn initially is used for both pooled and individual tests). The probability of a positive pooled blood sample is $1 - 0.99^{20} \approx 0.18$, so if people are divided into groups of 20, about 18% of the groups need to undergo the individual testing. One practical concern is that if groups are too large, the pooled blood sample might become too diluted and single individuals who are sick may go undetected. In the case of syphilis, however, Dorfman points out that the diagnostic test is extremely sensitive and will detect the antigen even in very small concentrations. Dorfman's original article, bearing the somewhat politically

[1]Epidemiologists use the term *prevalence* for the proportion of individuals with a certain disease or condition. For example, a prevalence of 25 in 1,000 for us translates into $p = 0.025$. A related term is *incidence*; this is the proportion of *new* cases in some specific time-period.

incorrect title “The detection of defective members of large populations” was published in 1943 in the *Annals of Mathematical Statistics*. The procedure of pooling has many applications other than blood tests, for example, tests of water, air, or soil quality.

Let me finish with a little treat for the theory buffs. First of all, the expected value is commonly denoted by $\mu$ (Greek letter “mu”). The general formula for $\mu$ is as follows. If the possible values are $x_1, x_2, ...$, and these occur with probabilities $p_1, p_2, ...$, respectively, the expected value is defined as

$$\mu = x_1 \times p_1 + x_2 \times p_2 + \cdots$$

where the summation goes on for as long as it is needed. In the case of a die roll, the summation stops after six terms, $x_k$ equals $k$ and all $p_k$ equal $1/6$. For another example, recall the binomial distribution from page **??**. This counts the number of successes in $n$ independent trials where each time the success probability is $p$. The possible outcomes are the numbers 0, 1, ..., $n$, and the corresponding probabilities were given in the formula on page **??**. The expected number of successes is therefore

$$\mu = \sum_{k=0}^{n} k \times \binom{n}{k} \times p^k \times (1-p)^{n-k}$$

which is not completely trivial to compute. However, it is easy to guess what it is. For example, if you toss a coin 100 times, what is the expected number of heads? Fifty. If you roll a die 600 times, what is the expected number of 6s? One hundred. In both cases, the expected number is the product of the number of trials and the success probability, and this is true in general. Thus, the binomial distribution with parameters $n$ and $p$ has expected value $n \times p$ (which as usual for expected values is not necessarily a possible actual outcome). If you are familiar with Newton’s binomial theorem, you might be able to show that the expression for $\mu$ above indeed equals $n \times p$.

## GOOD THINGS COME TO THOSE WHO WAIT

There are expected values where the summation in the formula from the previous section goes on forever. This does not mean that it takes forever to compute them, only that we can get an infinite sum if there is no obvious limit on the number of outcomes. For example, if you toss a coin repeatedly and count how many tosses it takes you to get heads for the first time, this number can theoretically be any positive integer. Although it is highly unlikely that you have to wait until the 643rd toss, you cannot rule it out. There is thus an infinite number of outcomes. I have already pointed out that probabilists do not fear the infinite, and our notation for the expected value in a case like this is

$$\mu = \sum_{k=1}^{\infty} x_k \times p_k$$

where $\infty$ is the infinity symbol, indicating that the sum never ends. It is one of the little intricacies of higher mathematics that you can add an infinite number of terms and still end up with a finite number. The probabilities $p_k$ must of course eventually become very, very small. The probability that you get your first head in the 643rd toss is, for example, $(1/2)^{643}$, which starts with 193 zeros after the decimal point. In general, the probability that you get the first head in the $k$th toss is $(1/2)^k$, and the expected number of tosses until you get heads is

$$\sum_{k=1}^{\infty} k \times (1/2)^k = 1 \times (1/2) + 2 \times (1/2)^2 + 3 \times (1/2)^3 + \cdots$$

and, believe it or not, this messy expression equals 2. This is intuitively appealing though. As heads show up on average half the time, they appear on average every other toss and your expected wait ought to be two tosses. By changing the success probability from $1/2$ to $1/6$, an even messier sum can be shown to equal six; thus, the expected wait until you roll a 6 with a die is six rolls. And yet another change, to $1/38$, reveals that each roulette number is expected to show up once every 38 spins. In general, if you are waiting for something that occurs with probability $p$, your expected wait is $1/p$. One of the rewards of studying probability is that mathematics and intuition often agree in this way. Another reward is of course that math and intuition do often *not* agree, at least not immediately, thus yielding wonderfully surprising results. As you have already learned, probability certainly has a complex and contradictory charm.

There is another way to compute the expected wait until the first heads, or 6, or roulette win, a way that avoids the infinite sum. Recall how we, starting on page **??**, computed winning probabilities in some racket sport problems by considering a few different cases, one of which led back to the starting point, thus giving an equation for the unknown probability. We can use such a *recursive* method here too. Suppose that you wait for something that occurs with probability $p$ and let $\mu$ denote the expected wait. In the first trial, you either get your event of interest or you do not. If you do, the wait was one trial. If you do not, you have spent one trial and start over with an additional expected wait of $\mu$ trials, yielding a total of $1 + \mu$ expected trials. As the first case has probability $p$ and the second $1 - p$, you get the following equation for $\mu$:

$$\begin{aligned}\mu &= p \times 1 + (1 - p) \times (1 + \mu) \\ &= 1 + \mu - p \times \mu\end{aligned}$$

which simplifies further to the equation $0 = 1 - p \times \mu$ that has solution $\mu = 1/p$, just like we wanted.

Let us look at a variant of the problem of waiting for something to happen. In the *Seinfeld* episode "The Doll," Jerry is very happy to find a dinosaur in a cereal box (right after Elaine has told him he is juvenile). Let us now say that there are ten different plastic toys to be found in that type of cereal box. In order to get all of them, what is the expected number of boxes Jerry must buy? This is difficult to solve directly by using the definition of expected value. In order to do this, you would have to compute the probability that it requires $k$ boxes for the possible values of $k$, and as there is no upper limit on these values, this presents a tricky problem. Just try to compute the probability that Jerry must buy 376 or 12,971 boxes.

We will do something smarter. First of all, one box is bought and contains a dinosaur. What is the expected number of boxes Jerry must buy in order to get a different toy? As the probability to get a different toy is $9/10$, he can expect to buy $10/9$ boxes, in analogy with what I said above with $p = 9/10$. Once he has gotten two different toys, he starts waiting for one different from these, and as there are now eight remaining toys, the probability to get something different is $8/10$ and the expected number of boxes is $10/8$. Next, he can expect to buy another $10/7$ boxes, then $10/6$, and so on until he finally can expect to buy $10/2 = 5$ boxes to get the second-to-last toy and $10/1 = 10$ boxes to get the final toy. Finally, in order to get the expected number of boxes Jerry must buy to get all the toys, we use the additivity property of expected values and conclude that he can expect to buy

$$1 + 10/9 + 10/8 + \cdots 10/2 + 10/1 \approx 29$$

boxes. Note that one third of these are bought in order to get the very last toy, every parent's nightmare. The expression above can be rewritten in a mathematically more attractive way as

$$10 \times \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{9} + \frac{1}{10}\right)$$

where the expression in parenthesis consists of the first 10 terms of the *harmonic series*. It is a well-known mathematical result that as more and more terms are added, the harmonic series summed up to $n$ terms, $H_n$, gets close to the natural logarithm of $n$, denoted $\log n$ (or sometimes $\ln n$). The natural logarithm of a number $x$ is what the number $e$ ($=$ 2.71828..., remember the discussion on page **??**) must be raised to in order to get $x$. Thus, if $e^y = x$, then $y$ is the natural logarithm of $x$: $y = \log x$.[2] As $n$ increases, the difference $H_n - \log n$ approaches a number that is known as *Euler's constant* and is approximately equal to 0.58.[3] We can now establish a nice general formula for the expected number of cereal boxes if there are $n$ different toys:

$$\text{expected number of boxes} \approx n \times (\log n + 0.58)$$

which for $n = 10$ gives 28.8, approximately 29 just like above. If $n$ is very large, we need to refine the constant 0.58; see footnote 3. This type of problem did not start with Jerry Seinfeld; it is a classic probability problem usually called the *coupon collecting problem* and has been generalized in a multitude of ways.

A related type of problem is the so-called *occupancy problem*. If Jerry learns that he can expect to buy 29 boxes in order to get all the toys and decides to go on a cereal shopping spree and buy 29 boxes at once, how many

[2]A more familiar logarithm is the base-10 logarithm where $e$ is replaced by 10. For example, the base-10 logarithm of 100 is 2 because this is what 10 must be raised to in order to get 100: $10^2 = 100$. In a similar way, you can consider the logarithm in any base and, for example, conclude that the base-4 logarithm of 64 is 3 because $4^3 = 64$. The ancient Babylonians liked the base 60 and we still use this to keep track of time in seconds, minutes, and hours. In our everyday math, we use base 10, computer scientists like the bases 2 (binary system) and 16 (hexadecimal system), but to mathematicians only the base $e$ is worthy of consideration.

[3]Leonhard Euler, a Swiss mathematician who lived between 1707 and 1783, was one of the greatest mathematicians ever. He was extremely prolific and contributed to almost every branch of mathematics. His collected works fill over 70 volumes, and his name has been given to so many mathematical results that when you refer to "Euler's Theorem" you have to specify *which* Euler's theorem that you are talking about. The constant mentioned here has an infinite decimal expansion starting with 0.5772156..., following no discernible pattern, and it is a famous unsolved problem whether it is *rational* (can be written as a fraction of two integers) or *irrational*.

different toys can he expect to get? Note that it is *not* ten. Of course he could be really unlucky and get dinosaurs in all of them, which has probability $(1/10)^{29}$. Multiply this probability by 10 and get the probability that only one type of toy (not necessarily a dinosaur) is represented. It is also possible to have 2, 3, ..., 9, or 10 different toys represented in the 29 boxes. The expected number of toys must be somewhere between 1 and 10, but it is tricky to compute directly. Again, additivity comes to our rescue and in this case in a really clever way.

Open all of the boxes, and first look for dinosaurs. If you find any, count "one" and otherwise "zero" (note that you count "one" if you find *at least* one dinosaur; you do *not* count the *number* of dinosaurs). Next, look for another type of toy, say, a SAAB 900 (a car model featured in several *Seinfeld* episodes). If you find any, count "one" and otherwise "zero." Keep looking for other types of toys, each time counting "one" if you find it and "zero" otherwise. When you have done this ten times, you have ten ones and zeros, and if you add them, you get the number of different toys that are represented (if your sum equals ten, they are all there). Thus, to get the final number, you added ten numbers and by additivity of expected values, to get the expected final number, you add ten expected values, each such expected value being computed from something that can be either one or zero. Also note that all these individual expected values are equal because there is no difference between the toys in regard to how likely they are to be in the box. What then is such an expected value?

In order to find it, we only need to figure out the probabilities of "one" and "zero." If the probability of 1 is $p$, the probability of 0 is $1 - p$ and the expected value is

$$0 \times (1 - p) + 1 \times p = p$$

and by adding ten such expected values, we realize that the expected number of different toys is $10 \times p$. To find $p$, note that we count "one" if there is at least one dinosaur. The ever useful Trick Number One tells us that the probability of this is one minus the probability of no dinosaurs, and we get

$$\text{P}(\text{at least one dinosaur}) = 1 - (9/10)^{29}$$

and finally

$$\text{expected number of different toys} = 10 \times (1 - (9/10)^{29}) \approx 9.5$$

so the juvenile Jerry is quite likely to get all his toys.

Let us summarize the coupon collecting problem and the occupancy problem in a general setting. There are $n$ different types of objects and you are attempting to acquire them one by one. The expected number of attempts until you get all of the $n$ objects is

$$n \times \sum_{k=1}^{n} (1/k) \approx n \times (\log n + 0.58)$$

and if you try $N$ times, the expected number of different objects that you get is

$$n \times \left(1 - \left(\frac{n-1}{n}\right)^N\right)$$

where you can notice that this number is very close to $n$ if $N$ is large, as you would expect.

The zeros and ones that you summed above are called *indicators* because they indicate whether a certain type of toy is present in the boxes. Resorting to indicators is a very useful technique to compute expected values, another example being the *matches* that we discussed on page **??**. For a quick reminder, if you write down the integers 1, 2, ..., $n$ in random order, the probability that there are no matches (no numbers left in their original position) is approximately 0.37 regardless of $n$. It is also possible to compute the probability of one match, two matches, and so on, and from this we could compute the expected number of matches. However, to find the expected number of matches, it is easier to use indicators. Simply go through the sequence and count one whenever there is a match and zero otherwise. Add the zeros and ones to get the number of matches. To get the expected number of matches, we only need to figure out the probability of a match in a particular position and multiply this by $n$, just like with the toys in cereal boxes above. This is easy. Focus on a particular position. As the numbers are rearranged at random, the probability that this position regains its original number is simply $1/n$ and the expected number of matches is therefore $n \times 1/n = 1$. Regardless of how many men leave their hats at the party, when hats are randomly returned, one man is expected to get his own hat back.

## EXPECT THE UNEXPECTED

In the previous chapters we have seen many examples where probability calculations lead to results that are surprising or counterintuitive. This is the case for expected values as well, and we will look at several examples. First, some random geometry.

Suppose that you create a random square by rolling a die to determine its sidelength. You then also compute the area, which is the square of the sidelength. The possible sidelengths are thus 1, 2,..., 6; the possible areas are 1, 4,..., 36; and each sidelength S corresponds to precisely one area A according to the equation $A = S^2$. Plain and simple. Let us now compute the expected sidelength and area. The expected sidelength is easy; we already know that this is 3.5. For the expected area, we can then square this value and get $3.5^2 = 12.25$. Or can we? Better be careful and do the formal calculation. As each sidelength has probability $1/6$ and corresponds to exactly one area, each area also has probability $1/6$ and we get the expected area

$$1 \times 1/6 + 4 \times 1/6 + \cdots + 36 \times 1/6 \approx 15.2$$

which is not at all 12.25. Apparently we cannot just square the expected sidelength to get the expected area. This becomes clearer if we think about long-term averages. For example, occurrences of sidelengths 1 are in the long run compensated for by sidelengths 6 and they average 3.5. However, when you compute the corresponding areas, sidelength 1 gives area 1 and sidelength 6 gives area 36; these areas average 18.5, which is not the square of 3.5. In the same way, sidelengths 2 and 5 average 3.5, but the corresponding areas 4 and 25 average 14.5. When all areas are averaged, in the long run, the average will settle around 15.2. Notice that this number is *higher* than the square of the expected sidelength. This is because areas grow faster than sidelengths; doubling the sidelength quadruples the area. So when you say that "the average square has sidelength 3.5 and area 15.2," it may sound absurd but of course you will never actually see the "average square."

Here is a simple game. You and a friend are asked to take out your wallets and count your cash. The only rule of the game is that whomever has more must give it to the other (and if you have exactly the same amount, nothing happens). Would you agree to play this game? You might argue: "I know how much money I have. If my opponent has less, I lose what I have and if he has more, I win more than what I have. There is no specific reason to believe that he is poorer or wealthier than I am, so this seems like a good deal. In fact, since I have just learned about expected values, let me try to compute my expected gain. My $x$ dollars can lead to either a loss of $x$ dollars or a gain of $y$ dollars, where $y > x$ and since a gain and a loss each have probability $1/2$, my expected gain is

$$(-x) \times 1/2 + y \times 1/2 = (y - x)/2$$

which is always a positive amount."

The math formula looks impressive, and you no longer hesitate but conclude that the game is in your favor, and you accept to play. However, when you see the smug look on your opponent's face, you suddenly realize that he has gone through similar calculations and come to the conclusion that the game is in *his* favor, so he is also eager to play. This makes you confused. How can the game be favorable to *both* of you?

The paradox stems from your implicit assumption that you are equally likely to win or lose, regardless of the amount in your wallet (that is where the probability $1/2$ comes from). Clearly this is not true. For example, if you have no money at all, you are almost certain to win unless your opponent is also broke. At least you cannot lose anything. If you have some, but very little money, you are quite likely to win, but if you have a lot of cash, chances are that your opponent has less and you lose. Remember, "either/or" is not the same as "50–50."

Let us look at a simple example. Suppose that you and your opponent simply flip a coin each to decide how much cash you have. Heads means you have \$1, tails that you have \$2. If you and your opponent flip the same, nothing happens. If you flip heads and he flips tails, you win \$1; if you flip tails and he flips heads, you lose \$1. As these two scenarios are equally likely, your expected gain is \$0 and the game is fair.

OK, that was easy. Let us make it a little more complicated and suppose instead that you and your opponent choose your cash amounts by each rolling a die. What is your expected gain? First, we can ignore all ties. Second, there is a certain inherent symmetry in that, for example, the outcome $(3,5)$ (your amount first) has the same probability as the outcome $(5,3)$. In the first case you win \$2; in the second, you lose \$2. In this fashion, each gain is canceled by an equally probable loss of the same size, and as you sum over all possible outcomes, you end up with \$0 and the game is again fair.

Now, people don't go around and toss coins or roll dice to decide how much cash they have. But these were only examples to illustrate that we can describe the amount of money in a wallet at some arbitrary time as generated by some random mechanism. There is an amount of uncertainty in numbers and sizes of cash withdrawals and cash payments, and in the end, it is reasonable to assume that there is a range of possible cash amounts to which we can ascribe probabilities. It is fairly easy to show (and even easier to believe) that the expected gain for each player is $0, regardless of what this range and these probabilities are, as long as they are the same for both players.

One of the first to describe the wallet paradox was Belgian mathematician Maurice Kraitchik in his 1942 book *Mathematical Recreations*, but with neckties instead of cash. I found it in Martin Gardner's 1982 book *Aha! Gotcha*, a collection of various mathematical puzzles. Mr. Gardner does not seem to have fully grasped the problem though. In his own words, "We[4] have been unable to find a way to make this clear in any simple manner" and points out that Kraitchik himself "is no help." But Mr. Gardner also remarks that the paradox perhaps arises because each player "wrongly assumes his chances of winning or losing are equal," and as I explained above, this is precisely the resolution to the paradox. As I mentioned in Chapter 2, Mr. Gardner pursued a lifelong devotion to educating the general public in mathematics, and considering this noble task, let us forgive him his somewhat indecisive treatment of the wallet paradox.

The wallet paradox was puzzling at first, but I think we managed to eventually set it straight. The next paradox is similarly mindboggling and not so easy to resolve. You are presented two envelopes and are told that one contains twice as much money as the other. You choose an envelope at random, open it, and note that it contains $100. You are now asked if you want to keep the money or switch and take what is in the other envelope. First, there does not seem to be anything to gain from switching, but then you start thinking. The other envelope contains $50 or $200, and since you chose randomly, it is equally likely to be either. Thus, by switching you either gain $100 or lose $50, and your expected gain is

[4]You may have noticed that mathematicians are very fond of the *pluralis majestatis*, a manner of expression traditionally reserved for royalty. Mark Twain proposed to extend the privilege to people with tapeworms; mathematicians seem to have added themselves to the list. Personally I believe this is because mathematicians are a very friendly and communal minded bunch who often feel that manipulating math formulas is a lonely business.

$$(-50) \times 1/2 + 100 \times 1/2 = 25$$

so it seems to be to your advantage to switch.

OK, so switch then, what is the problem? Well, there is nothing special with the amount $100, and the calculations can be repeated for any amount A that you find in the first envelope; in which case, the other envelope contains $A/2$ or $2\times A$ and your expected gain is

$$(-A)/2 \times 1/2 + 2 \times A \times 1/2 = A/4$$

dollars. Thus, it is always to your advantage to switch, so why even bother opening the first envelope? Just take it and immediately switch to the other. But why even bother taking the first? Just take the other envelope directly! But wait, then that envelope has become the first so shouldn't you then switch to the other, formerly first, envelope? But then you should take that envelope directly instead. But then...

Now that was really confusing. Something must be wrong but what? Let us try to do the experiment and see what happens. We get two envelopes, put two amounts of money in them, and start choosing, opening, and switching. What will happen? Naturally, you win as often as you lose in the long run, and the amount you win or lose is always the same. There are *two* envelopes and *two* amounts of money, but above we had *three* possible amounts floating around: $A/2$, A, and $2\times A$. Even though you may observe A dollars in your envelope and have no reason to believe more in either of the amounts $A/2$ and $2\times A$ in the other, it does not seem sensible to translate this into probabilities the way we did above. Once again, "either/or" is not necessarily the same as "50–50." In this case, it is actually either "0–100" or "100–0," you just do not know which.

A better description is that you are presented two envelopes that contain A and $2\times A$, respectively, for some amount A. If you choose at random, open and switch, you are equally as likely to gain $A as you are to to lose $A. The world makes sense again, and the envelope problem is not fun anymore.

## SIZE MATTERS (AND LENGTH, AND AGE)

Consider a randomly sampled family with children. On average equally as many boys are born as girls; therefore, such a family has, on average, equally as many sons as daughters. But this must mean that boys tend to have more sisters than brothers. For example, in a family with four children, the average composition is two sons and two daughters, and in such an average family, each boy has two sisters but only one brother. On the other hand, once a boy is born, the rest of the children should be born in the usual 50–50 proportions, which indicates that boys tend to have equally as many brothers and sisters. What is correct?

The second claim is correct. Boys to *not* tend to have more sisters than brothers. This may seem paradoxical at first, though. If you sample a boy at random and he has on average the same number of brothers as sisters, once you add him to the mix does this not indicate that there tend to be more boys than girls in the family? Yes indeed, but there is a twist. There is a difference between sampling a *family* and sampling a *boy*. Indeed, when you sample a boy, you are ruling out the families that have only girls, always selecting a family that has at least one son, and *such* a family *does* on average have more sons than daughters. For a simple illustration, consider only families with two children so that the equally likely gender combinations listed by birth order are GG, GB, BG, and BB. If a family is sampled at random, the probability that it has no sons is $1/4$, the probability that it has one son is $1/2$, and the probability that it has two sons is $1/4$. The expected number of sons in the family is therefore

$$0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4 = 1$$

but if a *boy* is sampled at random, the number of sons *in his family* (himself included) is equally likely to be one or two, the reason being that you are now choosing from the four Bs, two of which are paired with a G and the other two with another B. The expected number of sons is therefore

$$1 \times 1/2 + 2 \times 1/2 = 1.5$$

When the sampled boy is removed, the remaining expected 0.5 sons just means that his sibling is equally likely to be male or female. Thus, the average family has exactly one son who still manages to have on average half a brother (not a half-brother, mind you). But just like “average square” earlier, “average family” is not a precise concept unless we specify how the sampling is done. You may also think about it like this: Suppose that children from 1,000 families are gathered at a meeting. There will then be roughly the same number of boys and girls present. Suppose instead that 1,000 *boys* are gathered and that each has brought all his siblings. In the entire group, there will then tend to be more boys than girls present, but among the siblings of the selected boys, proportions are still 50–50. Boys do not tend to have more sisters than brothers; rather, they tend to belong to families that have more sons than daughters.

If you did not get this right the first time, you are in good company. Our constant companion Sir Francis Galton noticed in his 1869 book *Hereditary Genius* that British judges were all men and came from families that had on average five children. He erroneously concluded that the judges therefore had on average 2.5 sisters and 1.5 brothers. Thirty-five years later he realized his mistake and corrected it in an article with the intriguing title "Average number of kinsfolk in each degree" published in the journal *Nature* in 1904 (following an even more intriguingly entitled article, "The forest-pig of Central Africa" by zoologist Philip L. Sclater).

When we sample a boy or a British judge rather than a family, this is an example of *size-biased sampling*. Let us take a closer look at the two-children family. If a family is sampled at random and the number of boys counted, this number can be 0, 1, or 2, and the corresponding probabilities are $1/4$, $1/2$, and $1/4$. In the terminology from page **??**, the probability distribution on the set $\{0, 1, 2\}$ is $(1/4, 1/2, 1/4)$. Now instead sample a boy. The probability distribution on the same set is then instead $(0, 1/2, 1/2)$, and the interesting thing is that these new probabilities can be obtained by multiplying each of the first three probabilities by its corresponding outcome: $0 = 0 \times 1/4, 1/2 = 1 \times 1/2$, and $1/2 = 2 \times 1/4$. In other words, the probabilities changed proportional to size: 0 boys became 0 times as likely, 1 boy as likely as before, and 2 boys twice as likely. The new probability distribution is therefore called a *size-biased distribution*.

For another example, roll a die. The set of possible outcomes is then the set $\{1, 2, 3, 4, 5, 6\}$ where each outcome has probability $1/6$. Rather than rolling the die, you can think of this as choosing a face of the die at random. Now instead choose a face of the die by first choosing a *spot* at random, and then choosing the face that this spot is on. As there are $1 + 2 + \cdots + 6 = 21$ spots, the probability to get the face showing 1 is $1/21$, the probability to get the face showing 2 is $2/21$,..., and the probability to get the face showing 6 is $6/21$. The probability distribution on the same set $\{1, 2, 3, 4, 5, 6\}$ is now $(1/21, 2/21, ..., 6/21)$ instead of the distribution $(1/6, 1/6, ..., 1/6)$ we get when we choose a face at random. If we follow the idea in the previous example with the two-children family and multiply each outcome with its corresponding probability in the old distribution, we get $(1\times1/6, 2\times1/6, ..., 6\times1/6)$, that is,

$(1/6, 2/6, ..., 6/6)$. This set of numbers is not a proper probability distribution because the sum of the numbers is not equal to one. However, if each number is multiplied by $6/21$, we get precisely the new distribution when a spot is chosen at random. Again, the probabilities in the new distribution have changed by a factor proportional to size. The new size-biased probability of $k$ is the old probability $1/6$ multiplied by $6 \times k/21$.

There is more to be said. As $21/6 = 3.5$, which is the expected value of a die roll, the size-biased probability is in fact the old probability times the size of the outcome divided by the expected value, $1/6 \times k/3.5$. Let us look at this more formally. Denote the old probability of $k$ by $p_k$, the expected value by $\mu$, and the size-biased probability by $\widehat{p}_k$. We then have the relation

$$\widehat{p}_k = k \times p_k/\mu$$

for $k = 1, 2, ..., 6$. In our particular case, the $p_k$ are all equal to $1/6$ and $\mu = 3.5$, but the relation we stated between the $p_k$ and the $\widehat{p}_k$ is true for any probability distribution on any set. The size-biased distribution is the old distribution with each probability multiplied by $k/\mu$.

For another example of size-biased sampling, suppose that you choose a U.S. state by randomly sampling and recording the state of (a) a U.S. Senator and (b) a member of the U.S. House of Representatives. Then (a) is equivalent to choosing a state at random, whereas (b) is size-biased sampling because larger states have more House representatives and are thus more likely to be chosen. If you want all states to be equally likely, choosing a member of the House is incorrect, but if you want to give more weight to more populous states, it is correct. In general, size-biased sampling may be something you do not wish to do and that happens by mistake, but it may also be precisely what you want to do. There are many real-life situations where some type of size-bias becomes an issue. When an individual is chosen at random for an opinion poll, she is likely to come from a family that is larger than average, live in a city that is larger than average, go to a school that is larger than average, work for a company that is larger than average, and so on, all of these being factors that may have an impact on her opinions. When an ichthyologist catches fish, this may be done by detecting an entire school and larger schools are easier to detect. The same situation arises for any kind of animal that appears in clusters, be it flocks of birds, armies of frogs, or smacks of jellyfish. When a forest is inspected from the air for a tree disease, larger patches of sick trees are easier to detect. Larger tumors are easier to detect on a scan or X ray. And so on and so forth; size definitely matters.

Now let our randomly chosen family take a trip to Yellowstone National Park where the most visited attraction is the *Old Faithful* geyser, famed for its regular eruptions, which occur about every 90 minutes. When our friends arrive, they would thus expect to wait 45 minutes for an eruption. As they wait, they start talking to a man who has visited many times and has carefully recorded his waiting times, which average more than 45 minutes. He tells our family that this indicates that the geyser is slowing down, but data from the park rangers do not give such indications. Other than that our family's new friend may have had some bad luck, is there a logical explanation?

Definitely. The crux is that the Old Faithful, contrary to her name and reputation, does not erupt *exactly* every 90 minutes, only on average. Indeed, times between eruptions vary between 30 minutes and 2 hours but are most typically in the 60–100-minute range or so. If it did erupt exactly every 90 minutes and you arrived at a random time, your expected wait would certainly be 45 minutes. But now that intervals vary in length, you are in fact more likely to arrive in one of the longer intervals and thus your expected wait is longer than 45 minutes. To simplify things, suppose that intervals alternate between one and two hours so that eruptions occur at noon, 2 P.M., 3 P.M., 5 P.M., 6 P.M., and so on. The average interval length is then 90 minutes, but if you arrive at random, you are twice as likely to arrive in a 2-hour interval and your expected wait is one hour; if you arrive in a 1-hour interval, your expected wait is half an hour. Thus, two thirds of the time you wait on average an hour and one third of the time, half an hour. As $2/3 \times 1 + 1/3 \times 1/2 = 5/6$, your expected wait is $5/6$ of an hour or 50 minutes, longer than half the average interval time 45 minutes. See Figure 5.1 for an illustration of this scenario. In reality there is of course much more randomness than just shifting back and forth between one- and 2-hour intervals but you get the general picture.

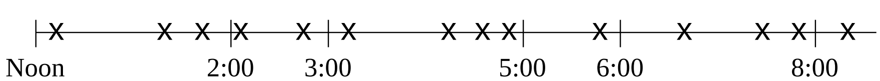


**Figure 5.1** The Old Faithful erupting at alternating intervals of lengths one hour and two hours and successive random arrivals. Note that there are more arrivals in the 2-hour intervals, making the average waiting time for an eruption more than 45 minutes.

The situation described above is an example of the *waiting time paradox*, a well-known phenomenon in probability. Another example of the waiting time paradox is when you catch a bus by randomly arriving at a bus stop. Even though the bus may run on average twice an hour, due to random variation, you are more likely to hit the longer intervals and must wait on average more than the 15 minutes' waiting time you would have if they ran exactly every half-hour. However, bus arrivals are still fairly regular and the difference is not likely to be large. It is not until the case of the rare and unpredictable events we studied in Chapter 3 that the name "paradox" is really earned. Let us look at earthquakes as an example. According to the U.S. Geological Survey, great earthquakes (magnitude 8 and higher on the Richter scale) occur on average once a year worldwide. Considering the capricious nature of earthquakes, let us agree that they qualify as rare and unpredictable. But this means that at any given time, the expected waiting time until the next great earthquake is one year, regardless of when the previous earthquake occurred, so if a space alien decides to pay a surprise visit to Earth, he can expect to wait one year for the next earthquake. On the other hand, when he arrives, the expected time since the *last* earthquake is also one year (just think of time running backward). One year since the last earthquake, one year until the next, yet one year between earthquakes and not two! Seems paradoxical but remember that these are expected values, and our alien friend is simply more likely to arrive in an interval that is longer than usual. Very short intervals that contribute to lowering the expected length are likely to be missed completely.

The waiting time paradox has a lot in common with size-biased sampling. Consider, for example, the simplified Old Faithful example with intervals between eruptions that are equally likely to be one hour or two hours. A randomly sampled interval is then equally likely to be of either length, and its expected length is 90 minutes. However, when you *arrive* at random, you can think of this as sampling an interval where the 2-hour interval is twice as likely as the 1-hour interval. Thus, the initial probability distribution $(1/2, 1/2)$ on the set $\{30,60\}$ (minutes) has changed to $(1/3, 2/3)$, where more weight is given to the larger value. Note how the new probabilities are proportional to the old probabilities times the interval lengths. Thus, the new distribution is size-biased or, more appropriately in this case, *length-biased*. This was the simplified example but regardless of what the real distribution of inter-eruption times are, when you arrive at random you choose such an interval with a probability proportional to its length.

A similar type of bias shows up when *life expectancy* is computed. In our terminology, life expectancy is the expected lifespan of a newborn individual. In a human population, life expectancy is estimated by recording the ages of everybody who dies (usually in a year) and taking the average. In the *Seinfeld* episode “The Shower Head,” George Costanza tries to convince his parents to move to Florida by pointing out that life expectancy in Florida is 81 and in Queens where they live, 73. Does this mean that Frank and Estelle could expect to live eight years longer in Florida? Not quite. One reason (other than the orange juice) that Florida has a high life expectancy is that many people move there from other states, most notably New York. As these people have already started their lives, and in most cases lived a good part of it, they cannot die at an age lower than that of their move. Thus, they “deprive” Florida of deaths at a young age, and this increases the average age at death. This scenario is typical for any city, state, or nation that has net immigration, another well-known (likewise orange cultivating) example being Israel. At the other end, states with net emigration have lower life expectancies. To help you understand, consider an extreme example and suppose that people born in A-town die either at age 40 or at age 80. They live and work in A-town, and if they survive age 40, they retire at 65 and then move to B-town where they live the rest of their lives. Life expectancy in A-town is 40 and in B-town 80, even though the people are really the same. Introduce a more realistic variability in lifespans and migration ages and you get a less drastic but similar effect.

## DEVIANT BEHAVIOR

Let us once again sit down at the roulette table.[5] Other than betting on a single number, there are plenty of ways to bet on a whole group of numbers. On the roulette table, the numbers 1–36 are laid out in a 3 by 12 grid where the top row is 1–2–3, the second row 4–5–6, and so on. Also, half of these numbers are red and half are black. On top of this grid are the numbers 0 and 00, colored green (on American roulette tables; European tables do not have the double zero). To bet on a single number is called a *straight bet*. You can also, for example, place an *odd bet*, which does not mean that you are betting in an unusual manner but that you win if any of the odd numbers 1, 3,..., 35 comes up. Likewise, you can bet on even or on red or black. You can

[5] I am constantly whetting your appetite with little glimpses into the world of gambling. Be patient. In Chapter 7, we will indulge shamelessly in all kinds of games, bets, and gambles.

also do *split bets*, *street bets*, *square bets*, *column bets*, and yet some. This is casino lingo, and all it means is that you can place your chip so that it marks more than one number and you then win if any of your numbers come up. Needless to say, the amount you win is smaller the more numbers you have chosen and the payouts are carefully calculated so that you lose 5 cents per $1 regardless of how you play. For example, let us say that you wager your dollar on an odd bet. The payout of such a bet is $1, and since there are 18 odd numbers between 1 and 36, the probability that you win $1 is $18/38$ and with probability $20/38$ you lose your wagered dollar. Your expected gain is therefore

$$1 \times 18/38 + (-1) \times 20/38 = -2/38 \approx -0.05$$

an expected loss of 5 cents per $1, just like if you place a straight bet. With the odd bet, your chances of winning are significantly higher than with the straight bet, but when you win, the payout is much smaller. In other words, the variability of your fortune is much greater when you place straight bets. This fact is not reflected in the expected value, so it would be nice to have a way to measure variability, in other words, to measure how much the *actual* value tends to differ from the *expected* value. There are different ways to do this, but probabilists and statisticians have come to the consensus that the best measure of variability is something called the *variance*. This is defined as the expected value of the square of the difference between the actual value and the expected value.[6] That was a mouthful. Let me illustrate it with the roulette example of odd bets. The expected value of your gain is $-0.05$ (dollars) and the two possible actual values are $-1$ and 1. The differences from the expected value are $-1 - (-0.05) = -0.95$ and $1 - (-0.05) = 1.05$ respectively. Square these two values to get $(-0.95)^2 = 0.9025$ and $1.05^2 = 1.1025$. Finally, we need to compute the expected value of these squared differences. As the first of them corresponds to a loss, it has probability $20/38$ and the second, corresponding to a win, has probability $18/38$. This gives the variance as the

[6]Squares are computed because we want to have only positive values. Another way to achieve this would be to compute *absolute values* of the differences between actual and expected values (i.e., the differences without signs). It turns out that squares have nicer mathematical properties than absolute values; for example, with some restrictions, variances are additive just like expected values, something that would not be true if we had used absolute values instead of squares.

expected value of these two squares:

$$0.9025 \times 20/38 + 1.1025 \times 18/38 \approx 1$$

a number that in itself does not mean much, but let us compare with the straight bet. Here, the possible actual values are $-1$ and 35 and a similar calculation to the one above gives a variance that is approximately 33. The much larger value of the variance of the gain of a straight bet than that of an odd bet reflects the larger variability in your fortune with the straight bet. In the long run, you lose just as much with either type of bet, but the paths to ruin look different.

The variance thus supplements the expected value in a useful way. Let us look at another example, the inexhaustible conversation topic of weather. Two U.S. cities that for different reasons caught my attention in early 2006 were Arcata and Detroit. In January 2006, I visited Arcata on the coast of northern California. Browsing through some weather statistics, I calculated that the daily high temperatures have an annual average of about 59 degrees Fahrenheit. A few weeks later, Super Bowl XL was played in Detroit, which has the same annual average daily high of about 59 degrees. Choose a day of the year at random to visit Arcata or Detroit, and the expected daily high is the same, 59 degrees. However, this does not mean much until it is also supplemented with the variance, which for Arcata is 12 and for Detroit 363 (and I challenge you to find a place with a lower temperature variance than Arcata). The much larger variance of Detroit reflects the larger variability in temperatures over the year. For example, the average daily high in Detroit in January is 33 and in July, 85. The corresponding numbers for Arcata are 55 and 63. In Detroit you will need to bring shorts or long johns depending on the season; in Arcata, none of these garments are of much use (but bring an umbrella in the winter).

I mentioned in passing above that there is no clear meaning of the value of the variance. One problem is that it is computed from values that have been squared, which means that the units of measurement have also been squared. What does it mean that the variance is 33 square dollars or 363 square degrees? Nothing, obviously, but there is an easy fix: Compute the square root of the variance. This number is called the *standard deviation* and is more meaningful because the unit of measurement is preserved. In the roulette example, the standard deviations for straight and odd bets are \$1 and $\sqrt{33} \approx$ \$5.7, respectively. In the weather example, the standard deviation for Arcata is 3.5 degrees and for Detroit, 19 degrees.

This feels a little better, but the standard deviation still does not have the crystal clear interpretation that the expected value has. There are some rules and results that can help, one of them due to another great Russian mathematician, Pafnuty Lvovich Chebyshev, who lived between 1821 and 1894 and is famous for his contributions to probability, analysis, mechanics, and, above all, number theory.[7] His result, known as *Chebyshev's inequality*, states that in any experiment, the probability to get an outcome within $k$ standard deviations of the expected value is at least $1 - 1/k^2$, for any value of $k$. For example, choosing $k = 2$ informs us that regardless of what the experiment is, the probability to get an outcome within two standard deviations of the expected value is at least 0.75. Stated differently, Chebyshev's inequality tells us that in a set of observations, at least 75% of the observations fall within two standard deviations of the average. In Arcata, we can expect at least 273 days with a daily high temperature between 52 and 66, and in Detroit, we can expect at least 273 days between 21 and 97 degrees. And with $k = 3$, we get $1 - 1/k^2 = 8/9 \approx 0.89$; at least 89% of observations are within three standard deviations of the expected value.

I would like to stress the "at least" part of Chebyshev's inequality. In reality the probabilities and percentages are often significantly higher. For example, in the roulette example with odd bets, *all* observations are within two standard deviations. Also note that if you choose $k = 1$, all Chebyshev tells you is that at least 0% of the observations are within one standard deviation. Certainly true but not very helpful. Chebyshev's inequality tends to be crude in this way but that is only natural because it is always true, regardless of the particulars of the experiment. It is sort of like saying that every U.S. state is smaller than 572,000 square miles in area. This is needed to include Alaska and is certainly true if we only consider the continental United States, but then 262,000 square miles would be enough. And if we restrict ourselves to New England, even less is needed. Despite these shortcomings, Chebyshev's inequality is still useful as we will learn in the next section.

Let me again pander to those of you who suffer from theory cravings and give the formal definition of variance. Suppose that our experiment can result in the outcomes $x_1, x_2, ...$, and that these occur with probabilities $p_1, p_2, ...$,

[7] Chebyshev also holds the unofficial world record among mathematicians for most spellings of last name. I should really say transliterations rather than spellings because in his native Cyrillic alphabet he is Чебышёв and nothing else. In the Western world, he has appeared in print in about a dozen different forms ranging from the minimalist Spanish version *Cebysev* to the consonant-indulgence of the German *Tschebyscheff*.

the same setup as we had when we formally defined expected value earlier. Denote this expected value by $\mu$, and remember that the variance involves computing the squared differences between each possible value and $\mu$, then computing the expected value of these squared differences. Translating this verbal description into mathematics gives the formal definition of the variance, commonly denoted by the symbol $\sigma^2$ (square of the Greek letter "sigma") as

$$\sigma^2 = (x_1 - \mu)^2 \times p_1 + (x_2 - \mu)^2 \times p_2 + \cdots$$

where the summation stops eventually if there are a finite number of outcomes and goes on forever otherwise. Check for yourself that this is precisely what we did above in the roulette examples. Just for practice, let us do the variance for the roll of a die. The possible values are 1, 2,..., 6, each with probability $1/6$, and the expected value is 3.5. The variance is therefore

$$(1 - 3.5)^2 \times 1/6 + (2 - 3.5)^2 \times 1/6 + \cdots + (6 - 3.5)^2 \times 1/6 \approx 2.9$$

which gives a standard deviation of 1.7. Let us compare this with the standard deviation of a die that has 1 on three sides and 6 on the remaining three. This die gives 1 or 6, each with probability $1/2$, so it also has an expected value 3.5. Its variance is

$$(1 - 3.5)^2 \times 1/2 + (6 - 3.5)^2 \times 1/2 = 6.25$$

which gives standard deviation 2.5. This is larger than the standard deviation of the ordinary die because this special die has outcomes that tend to be further away from the expected value 3.5. Again, we have an example where the expected value does not tell the full story but is nicely supplemented by the standard deviation.

Recall that the standard deviation is the square root of the variance, and it is therefore denoted by $\sigma$ and we can state the formal version of Chebyshev's inequality. Before we do that, though, let me mention an important concept in probability. Before any experiment, the outcome is unknown and we can denote it by $X$, which means that $X$ is unknown before the experiment and gets a numerical value after. Such an unknown quantity whose value is determined by the randomness of some experiment is called a *random variable*. This is a very important concept in probability that greatly simplifies the notation in many examples. If a die is rolled, instead of writing things like "the probability to get 5" and "the probability to get 6," we can first denote the outcome of the die by $X$ and write $\mathrm{P}(X = 5)$ and $\mathrm{P}(X = 6)$, a mathematical and more convenient notation. Chebyshev's inequality can now be stated as

$$\mathrm{P}(\mu - k \times \sigma \leq X \leq \mu + k \times \sigma) \geq 1 - 1/k^2$$

or, using absolute values,

$$\mathrm{P}(|X - \mu| \leq k \times \sigma) \geq 1 - 1/k^2$$

for any value of $k$ (which by the way does not have to be an integer; it could be 1.5 or 4.26 or any other nonnegative number). Make sure that these last two expressions are equivalent, and that they agree with the verbal description of Chebyshev's inequality that I gave earlier.

## FINAL WORD

The concept of expected value that we have investigated in this chapter can be thought of as the ideal average in a random experiment. The expected value summarizes the experiment in a single number, but we have seen many examples of how some care must be taken in the interpretation of this. The expected value's constant companion is the standard deviation that measures the amount of variability in the experiment, and together the two, $\mu$ and $\sigma$, provide a convenient summary of the random experiment. I have also at times hinted that we can interpret the expected value as the long-term average, and in the next chapter, this particular interpretation will be thoroughly investigated.

# 6

# Inevitable Probabilities: Two Fascinating Mathematical Results

# 7

# Gambling Probabilities: Why Donald Trump Is Richer than You

# 8

# Guessing Probabilities: Enter the Statisticians

# 9

# Faking Probabilities: Computer Simulation

# Index