Thinking About Uncertainty: An Introduction to Probability and Statistics

David Draper
University of California, Santa Cruz

©2004 David Draper

All rights reserved. No part of this book may be reprinted, reproduced, or utilized in any form or by any electronic, mechanical or other means, now known or hereafter invented, including photocopying and recording, or by an information storage or retrieval system, without permission in writing from the author.
Chapter 9

Expected Value, Standard Error, and the Central Limit Theorem

This chapter is about how to build probability models that help answer questions about random events, in science and daily life, that are hard to answer with the methods of Chapters 7 and 8. As the case study that will help indicate how to do this, I want to analyze a gambling game called Keno that people play in places like Las Vegas and Atlantic City. As we saw back in Chapter 7, gambling is a fitting choice historically for learning about probability, and if you have ever thought about trying your hand at the casinos with the goal of winning money I hope this chapter will dissuade you—it will become clear that Keno, like virtually all casino gambling games, is a losing proposition.

Case Study 9: Keno. In the casino game of Keno there is a big cylinder with balls numbered from 1 to 80 in it, and every now and then throughout the day somebody shakes up the balls in the cylinder and draws out 20 of them at random without replacement. For awhile before this drawing takes place, other people have been going around the casino offering the gamblers a chance to bet on which balls will be chosen.

Two available bets in Keno are a single number and a double
number. With the single number bet you pick a number that you think is going to be among those chosen, circle it on one of the Keno cards that the casino management has thoughtfully placed near you, and give the Keno person your wager. If you bet $1 on a single number and it turns out to be one of the 20 chosen, the casino gives you your dollar back plus $2 in winnings—in casino language, the payoff is 2 to 1. If your number is not chosen, they keep your dollar and encourage you to try again: “Maybe your luck will be better next time.” A double number is just like a single number except that you circle two numbers on your Keno card, and if they are both among the 20 numbers chosen the casino pays you 11 to 1.

Is Keno a good game to play? Well, it depends on what you mean by “good”—for example, if it gives you pleasure to play it, that ought to count for something (this is an example of a decision-making concept called utility, which we will talk about in Chapter xx). Right now I want to focus on a different issue: do you think you will win if you play? Notice that if you were to play Keno several times, we are uncertain how each play will come out—sometimes you will win, sometimes you will lose—but notice also that the “at random without replacement” method of picking the winning numbers means that we ought to be able to use probability to quantify our uncertainty about the outcomes. So this makes me interested in the answers to questions like the following:

- What is your chance of winning on any one play?
- If you bet $1 each time, how much money do you expect to win or lose on any single play, on average? Give or take how much?
- If you had the patience to play a large number of times, say 100, betting $1 each time, how much money would you expect to have won or lost after the 100 bets? Give or take how much? What is your chance of coming out
ahead after 100 plays?

The answers to these questions may be different for the single- and double-number betting strategies. I'll lay out the story for single-number gambling first, and then use the machinery to address the same questions if you were to bet on double numbers instead. The argument is rather lengthy, and mathier than most of the rest of the book, but hanging on all the way through this chapter has a big payoff—the model presented here is the cornerstone for all the rest of the work that follows.

9.1 Building a Probability Model

Figure 1 is where we're headed: it's a probability model for single-number $1 betting at Keno. Probability models turn out to have three main ingredients, corresponding to the left, middle, and right parts of the figure: the population, the sample, and the imaginary dataset. Each of these elements of the model can be thought of as a dataset, with (as usual) rows for replicates and columns for variables. In this part of the course we will only be looking at one variable at a time, so for now each of these datasets will have only one column, but toward the end of the book we will look at situations with two or more variables. The main setup I'm going to describe here is frequentist in character, but we will see later that it's useful in Bayesian calculations too.

Diagrams like Figure 1 looks complicated, and when you first start to do them they take awhile to finish, but I recommend writing them out in full whenever you have the patience to do so. They enforce a useful discipline in reminding us of all of the judgment calls we need to make when trying to model the real world in probability terms, and—if you do enough problems like the ones in this chapter—eventually a problem will come along in which failure to attend to all the ingredients in the diagram will cause your model to be unrealistic in some important way. The terminology imaginary dataset used here is nonstandard; the official description of it would involve words like "simulating the distribution of a random variable," but I like "imaginary dataset" better. I have put some more information about random variables in the math interludes below, if you get interested.
9.1.1 The Population

Basic to the construction of these models is the idea that you are studying something that could come out differently each time you do it, and you are imagining doing it over and over. The population dataset records how things might come out, as measured by some variable that’s particularly interesting to you, each time you repeat the process you’re imagining repeating. In single-number wagering at Keno, for instance, we are imagining betting $1 on your favorite number over and over, and the variable we would like to keep track of is how much money changes hands on each play—what we might call “your net gain” on any single play, if we considered it from your point of view rather than from the point of view of the casino. Sometimes it’s convenient to have a symbol for the number of rows in the population dataset—let’s call it $N$.

To finish the population dataset I guess I should put some data in it. There are two things to pay attention to: the possible values the variable of interest could take on, and the probabilities of each of those possible values actually occurring. Here I said I was interested in how much money might change hands on any single play, from your point of view. With the single number strategy if you bet $1 on your favorite number and it loses, you lose your dollar—in other words, your net gain is $-1$. If your number is chosen you win $2—in other words, your net gain is $+2$.

So far so good. Now what are the chances associated with each of these outcomes? If there are 80 numbers and the Keno person only picks 20 of them, the chance of your favorite number being in there somewhere doesn’t sound very high. It’s actually not too hard to work out this probability, so let’s take a minute and do so.

Drawing 20 numbers at random without replacement from the list 1, 2, ..., 80 is like writing all the numbers down on cards, one number per card, shuffling the deck of 80 cards thoroughly, so that all possible rearrangements are equally likely, and then dealing out the top 20 cards. Consider the deck of cards just before you deal out the top 20, from the point of view of your lucky number—the single number you want to bet on in Keno. There are 80 different positions this number could occupy in the deck—for example, on top, second from the top, third from the top, and so on down to the bottom—and because you
shuffled thoroughly they’re all equally likely. Twenty of them are favorable to you winning the $1 bet—any of the first 20 positions will get your number chosen—so by the equally-likely model the chance of you winning must be \( \frac{20}{80} = \frac{1}{4} = 25\% \), and your chance of losing must be \( \frac{3}{4} = 75\% \). We have answered the first question in the case study: your chance of winning on any one play is 25%. Single-number betting in Keno is not a very good game when viewed in this way—you probably would have wanted the odds a little closer to even money than that.

We’re not quite done with this part of the model: we still have to set up the population dataset to reflect the fact that you have 1 chance in 4 of winning $2 and 3 chances in 4 of losing $1. The simplest way to do this is to put four numbers in the population: three −$1’s and one +$2. (It would be equally good for what we’re going to do to put six −$1’s and two +$2’s in there, or any other choice that gives three times as many −$1’s as +$2’s, but three −$1’s and one +$2 is easiest.) Notice in the left-hand part of Figure 1 that it is helpful to other people trying to understand your model to put a box over the population dataset and fill it with words describing the population. In this case the population is hypothetical, or conceptual—it represents what might happen each time you make one $1 bet on a single number. In the statistical models we will build in Chapter xx, which are closely related to the probability models of this chapter, the population dataset will often be less hypothetical than it is here.

The last thing to specify about the population dataset is the same kind of summaries for it that we have been using to summarize a column of numbers ever since Chapter 3—center, spread, and histogram shape. In probability models like those we will look at, people always use the mean and SD to keep track of center and spread (we will see why a little later), and the best way to convey the histogram shape is to actually sketch it. So you finish off the population part of the probability model by writing the mean and SD of the numbers in the population right below the dataset itself (as in Figure 1), and below that you draw the histogram. In this case the histogram is discrete, with 3/4 of its area in a bar at −$1 and the other 1/4 in a bar at +$2.

It will turn out below that other parts of the model diagram will also have means and SDs, so we need some special notation for the population mean and SD to keep from getting confused. By convention
everybody gives Greek letters to population summaries, usually chosen to remind you of the first letter of the thing they summarize, so the population mean is usually called $\mu$ ("mu," pronounced "mew") and the population SD is typically called $\sigma$ ("sigma").

**Working Out the Population Mean and SD.** How do you calculate $\mu$ and $\sigma$? It's pretty easy—you just think of the numbers in the population as data, and figure out their mean and SD, much as we did in Chapter 3. The mean is particularly easy:

$$\mu = \frac{3(-\$1) + 1(+$2)}{4} = \frac{-\$1}{4} = -\$0.25.$$  \hfill (9.1)

This answers part of the second question in the case study: if you played $\$1$ single-number Keno a lot, in the long run 75% of the time you'd lose a buck and 25% of the time you'd win $\$2, which averages out to losing a quarter each time you play. People summarize this by saying that the **expected value** of your net gain on any given $\$1$ play at single-number Keno is $\mu = -\$0.25$. You could think of this as the entertainment value of $\$1$ wagered in this way—if you don’t get 25% worth of fun out of watching yourself lose $\$1$ most of the time, maybe you shouldn’t play this game. We shouldn’t be surprised that this expected value came out negative—after all, the casinos are in business to make a profit, not to let you make one at their expense. If they had been sporting about it they could have set the payoff so that the expected value was $\$0$—people refer to games with $\$0$ expected values as *fair*—but that would have left them with a 50/50 chance of losing money on every gambler who walked in the door, which would be a pretty uncertain business proposition (see Problem 1).

**Math Interlude.** Before we go on it's useful to rewrite the calculation for the mean in a different, more symbolic way. Let’s let $a_1, a_2, \ldots, a_k$ stand for the possible values for the outcome of any single $\$1$ play and $p_1, p_2, \ldots, p_k$ stand for the probabilities of those values occurring, where $k$ is how many such values there are. Here $k = 2$ and the two possibilities are $a_1 = -\$1$ and $a_2 = +\$2$, with probabilities $p_1 = \frac{3}{4}$ and $p_2 = \frac{1}{4}$. Then we could rearrange equation (9.1) above to read $\mu = \frac{3(-\$1) + 1(+$2)}{4}$ =
\[ \frac{3}{4}(-$1) + \frac{1}{4}(+$2) = p_1a_1 + p_2a_2 \text{ and write the population mean } \\
\mu \text{ symbolically as} \]

\[ \mu = p_1a_1 + p_2a_2 + \ldots + p_k a_k = \sum_{i=1}^{k} p_i a_i . \quad (9.2) \]

You can see from this that \( \mu \) is a kind of weighted average of the possible values for each $1 gamble, with the weights given by the probabilities of those values occurring.

What about the SD \( \sigma \)? The story is similar to calculating an SD with data you are thinking of as a sample, as we did in Chapter 3:

\[ \sigma = \sqrt{\frac{3[-$1 - (-$0.25)]^2 + 1[+$2 - (-$0.25)]^2}{4}} = \sqrt{\frac{6.75}{4}} = $1.30. \quad (9.3) \]

This should seem familiar, except that with samples of data in Chapter 3 we used to divide not by (the number of values) but (the number of values \( - 1 \)). That’s the only difference between sample and population SDs—when it’s the population SD people are after, they divide by \( N \), not \( (N - 1) \). One informal way to understand why is to recall the idea of degrees of freedom from Chapter 3: when you’re working with a sample of data and you don’t know the population mean, you need to use the data to guess it in calculating the SD, which “burns up one degree of freedom,” so that instead of dividing by the number \( n \) of data values in calculating the sample SD you divide by \( n - 1 \); but in figuring out the population SD the population mean is known, and there is no need to lose a degree of freedom in guessing at it.

\textit{Math Interlude.} A symbolic expression for \( \sigma \) is also useful, and is easy to get by rewriting equation (9.3) above as

\[ \sigma = \sqrt{\frac{3}{4}[-$1 - (-$0.25)]^2 + \frac{1}{4}[+$2 - (-$0.25)]^2} , \quad (9.4) \]

which in terms of the \( a \)'s and \( p \)'s and \( \mu \) is just

\[ \sigma = \sqrt{p_1(a_1 - \mu)^2 + \ldots + p_k(a_k - \mu)^2} = \sqrt{\sum_{i=1}^{k} p_i(a_i - \mu)^2} . \quad (9.5) \]

7
So $\sigma^2$ is also a kind of weighted average, in this case of the squared deviations of the possible values from the mean $\mu$. In parallel with the terminology of Chapter 3 people call $\sigma^2$ the population variance since it's the square of the SD.

Equation (9.3) answers the rest of the second question in the case study: Given that we have already worked out that $\mu = -\$0.25$, the interpretation of $\sigma = \$1.30$ is that each time you wager $\$1 on a single number at Keno you expect to lose about a quarter, give or take about $\$1.30$. This only roughly describes what will actually happen on any given play in this case, since the only possible values for your net gain are $-$\$1 and $+$\$2, but even with only two possible values $\mu$ and $\sigma$ give you some idea of what to expect on each play, and we will see in Chapter 10 that give-or-take statements like this become more and more accurate as the number of possible values in the population grows.

9.1.2 The Sample

Okay, so much for the population part of the probability model, which helps you visualize what might happen if you made a single $\$1 gamble at Keno. The next part of the model is the sample, which helps simulate a bunch of such $\$1 plays at a time. The idea is that the sample represents one possible realization of the overall random process you're studying. Here, for instance, I want to simulate making 100 $\$1 plays on a single number, so I'll let the first number in the sample stand for the outcome of the first $\$1 play, the second number for the outcome of the second play, and so on up to 100 plays. This means that the sample can be thought of as a dataset with one column/variable (representing your net gain on each play) and 100 rows/replications, so that the sample size is $n = 100$. The middle part of Figure 1 shows the sample part of the model diagram—I have made up some hypothetical outcomes of each $\$1 bet, about 3/4 of which should be $-$\$1 and 1/4 of which should come out $+$\$2.

The next thing to specify about the sample dataset is its relationship to the population—in other words, to be specific about the sampling mechanism. In Chapter 6 we talked about two kinds of samples—probability samples, and samples of convenience. The random way
in which the balls are drawn at Keno makes the sample in Figure 1 a probability sample, which (you will remember from Chapter 6) is fortunate for us because it's much harder to draw valid conclusions from samples of convenience than it is from probability samples. But what kind of probability sample corresponds to repeated $1 betting at Keno?

In Chapter 7 we looked at the two simplest probability sampling mechanisms, in which the draws from the population are either made

- At random with replacement (IID sampling, we called it, which stood for independent identically distributed draws), or

- At random without replacement (which we called simple random sampling, or SRS).

Does either of those mechanisms describe repeated betting at Keno? Well, it can't be SRS, because if you made the draws without replacement and the population only had 4 elements in it (as in Figure 1) you'd run out a long time before you got to 100 draws, so maybe it's IID. Let's see: are the individual $1 plays at Keno independent of each other? Sure—each time they draw the 20 balls, they put them back in and mix them up thoroughly before the next game, so that what happens in one Keno game can't affect other games. And are the amounts you win or lose on the individual $1 plays identically distributed, by which I mean that they have the same possible values and probabilities of taking on those values? Again, yes—each Keno game is played out, in the sense of probability, in exactly the same way. So I guess the sampling mechanism in repeated betting at Keno is IID. It's useful to add this information to the model diagram by drawing an arrow from the population to the sample, with the notation "IID" above the arrow to explicitly show anybody looking at the diagram how the sample relates to the population. We will see in later chapters that important ingredients in the process of computing things like the probability of coming out ahead in Keno depend on the sampling mechanism.

Math Interlude. Sometimes it's useful to have symbols that stand for the individual draws from the population and the values they take on. The standard symbol for outcome variables
is \( Y \), and that's as good as any other choice here, but to keep from getting confused we'd better add a subscript for which draw we're thinking of, so let's let \( Y_1 \) stand for the first draw, \( Y_2 \) for the second, and so on down to \( Y_{100} = Y_n \) for the last draw, with \( Y_i \) standing for the generic \( i \)th draw. When I say that \( Y_2 \) stands for the second draw I mean that quite literally: \( Y_2 \) represents the process of making the second draw at random with replacement from the population in Figure 1. Formally \( Y_1, \ldots, Y_n \) are called random variables, which is a pretty good name for them: if you were to repeat each draw over and over you would notice that the outcomes vary randomly.

When people are being especially careful they draw a notational distinction between random variables and the values they take on, by letting \( Y_i \) stand for the process of drawing for the \( i \)th time from the population and \( y_i \) for the outcome you actually get when you make that draw. For example, with the simulated data I presented in the middle part of Figure 1 the second Keno game resulted in the loss of a dollar, so on that occasion \( y_2 \) came out \(-\$1\). With this notation you can talk about the probability that \( Y_2 \) is \(-\$1\)—symbolically, \( P(Y_2 = -\$1) \)—but it doesn't make sense to talk about the probability that \( y_2 \) is \(-\$1\): it either is or it isn't.

If you go on to take another probability or statistics course, the book you use might well summarize a setup like the population-and-sampling story for 100 single-number \$1 Keno bets with notation something like this:

\[
Y_i \text{ IID, } i = 1, \ldots, n = 100, \quad P(Y_i = y_i) = \begin{cases} 
\frac{3}{4} & \text{if } y_i = -\$1, \\
\frac{1}{4} & \text{if } y_i = +\$2, \\
0 & \text{otherwise.}
\end{cases}
\]

In this book we will use model diagrams like Figure 1 instead of this notation, but the diagram and this equation stand for the same set of assumptions about how Keno works.

The idea of random variables is slightly strange at first; don't be dismayed if it takes you awhile to get the hang of them. As I said above, they stand for the process of doing something “at random,” like playing Keno or some other gambling game, and—since we're not actually playing Keno, we're just imagining
what might happen if you were to play—they’re hypothetical, with no objective reality. In other words, random variables are a mathematical fiction, but they turn out to be a useful fiction.

Data—what you get in the sample part of Figure 1 when you actually play Keno, or at least simulate doing so—is the only part of the story told by the figure that has any objective reality. You will never see a random variable, but—if you work in a quantitative field—you will routinely see data, and you will often find it useful to think of your data as like the outcome of a random process. To express this sort of thing you hear people speaking technically make statements like “I am thinking of my data as realizations of IID random variables,” which would probably have sounded incomprehensible before you read this, but all that people mean by such a statement is that they have a diagram like Figure 1 in mind.

Next in specifying the sample part of the diagram is how to summarize the sample dataset. Here I’m not so much interested in the 100 actual outcomes as I am in what they imply about how much money I’ve won or lost overall, so I’d like to summarize the 100 data values with a number that keeps track of my net gain or loss at the end of the 100 gambles. If you think about it a minute, you will see that the sum of the 100 draws is just what we’re looking for. (For instance, if there were only two $1 plays and you lost both times, the sample values would both be $1 and their sum would be $2, which is just another way of saying correctly that you would be behind $2.) We’re not sure how the sum will come out—each time you had the patience to play 100 times the sum would very likely be different—so let’s just give it a symbol, S, which it’s useful to write at the bottom of the sample dataset as in Figure 1. Symbolically, in terms of S, answering the last question in the case study about your chance of coming out ahead amounts to computing $P(S > 0)$.

The last ingredient in the sample part of the probability model is a histogram of the 100 sampled values, which can be sketched below the sample as in the middle part of Figure 1. Since we’re uncertain about how each $1 gamble will come out, the histogram is uncertain, too, but it helps to think about how it would probably look. Well, each draw has a 75% chance to be a −$1 and a 25% chance to be a
+$2, so in the 100 draws there should be around three -$1’s for every +$2—in other words, the sample histogram should look a lot like the population histogram. That’s another way to say that IID sampling should produce a pretty representative sample, which we have already seen back in Chapter 6 makes good intuitive sense.

9.1.3 The Imaginary Dataset

Okay, we’ve got the population and the sample, we’ve figured out what it is about the sample that’s of principal interest—the sum $S$ of the 100 sampled values—and we’re trying to compute the probability that $S$ would come out greater than $0$ if you were to make 100 IID draws tomorrow. From Chapters 7 and 8 we have two main approaches to quantifying probabilities: frequentist and Bayesian. In problems like this the Bayesian story doesn’t add anything to the frequentist approach: there doesn’t seem to be any prior information lying around in this situation, and if we imagined betting with someone about whether $S$ will be positive or not, it’s not at all clear what odds we should give or take to make the bet fair (except to say that $P(S > 0)$ should be less than $50\%$ or else the casino is crazy). So let’s try the frequentist approach, which involves repeating over and over the process that leads to a value of $S$—that is, repeatedly making 100 $1\$ single-number Keno bets—and computing the percentage of time that your net gain $S$ comes out positive.

Recall from Chapter 7 that there are two ways to carry out this relative-frequency plan in practice: we can either actually do the repetitions (or something equivalent to them), which is called the simulation approach, or we can imagine doing them and then use math to work out the relative frequency we want. The math approach is preferable, when feasible, mainly because using math to figure out how to compute probabilities in one kind of problem often indicates how to compute them in other problems as well (whereas simulation tends to be less generalizable). The simulation approach has a long and honorable tradition—from the early days in which people were interested in probability (the mid 1600s) to the present, people have used simulation to approximate probabilities that are difficult to compute any other way—and the only difference between the simulation and math
approaches is whether you actually do the repetitions or just think about doing them, so it is useful to look first at the simulation story.

The Simulation Approach. Before the advent of computers in the 1950s, people used one version or another of what might be called physical or empirical simulation. With this approach in the case of Keno, for instance, you would

- Get one of those cylinders with the 80 balls in it that I mentioned at the beginning of the chapter, or create something similar to it (like a hat with slips of paper numbered from 1 to 80 in it),

- Draw 20 numbers at random without replacement and see if your favorite number was among the 20 chosen values, recording −$1 or +$2 accordingly,

- Do the previous step a total of 100 times and add the −$1’s and +$2’s to get a value of S, and

- Repeat the last two steps M times for some (preferably large) value of M, and work out the percentage of repetitions in which S came out positive.

This would take a lot of time and effort, and at the end of it all you would have would be an estimate of the probability we want, because in practice—out of boredom or fatigue—I bet you would settle for a relatively small value of M, and (as I mentioned in Chapter 7) the official frequentist value of \( P(S > 0) \) is only obtained by letting M go to infinity.

These days to carry out the simulation approach people would write a computer program instead of drawing slips of paper out of a hat, which would improve on the physical approach in two ways: it would be faster, and you could get an estimate of arbitrary accuracy just by running your computer long enough to make M really big. However, before we settle for simulation let’s see what’s involved in the math approach—as I said earlier, the probability model we’re building turns out to generalize pretty smoothly to almost all the rest of the problems we’re going to encounter in the rest of the book, so the effort to finish building it seems worthwhile.
**The Math Approach.** The math story goes just like the simulation approach I itemized a minute ago, except that you just imagine doing each step instead of actually doing it. The fundamental operation here is making 100 $1 single-number Keno bets and recording your net gain at the end of the 100 gambles, which we modeled in Figure 1 by (a) making 100 IID draws from the population to create the sample and (b) computing the sum $S$ of the 100 draws. To finish calculating $P(S > 80)$ we evidently need to imagine taking another IID sample of size $n = 100$ and computing the sum of these 100 draws, taking another IID sample and computing the sum of *those* draws, and so on, over and over. It is convenient to collect the values of $S$ you would get by doing this into a third dataset that might be called the *imaginary dataset*, as in the right-hand part of Figure 1. The first entry in the imaginary dataset is the sum $S$ based on the first sample, which I have said in the middle of Figure 1 might come out $-31$; the second entry is the value of $S$ based on the second sample of 100 IID draws, which might come out $-10$ (say); and so on. I have labeled this column of numbers "Possible Sums" in the figure to remind us of what the imaginary dataset is.

The number of rows in the imaginary dataset is just the number of times we imagine getting another value of $S$, which in the simulation approach I called $M$. I said above that to get the official value of $P(S > 80)$ you have to let $M$ go to infinity, so apparently the imaginary dataset should have an infinite number of rows. You can see that building these probability models would have been second-nature to [the Red Queen] in *Alice in Wonderland*, who ["always imagined two or three impossible things before breakfast."] I guess while we're imagining the rest of this, it doesn't hurt much more to imagine a dataset with an infinite number of values in it.

### 9.2 The Expected Value and Standard Error of a Sum

Okay, so we've got this imaginary dataset with a lot of values of $S$ in it—how do you figure out what percentage of these values are greater than 80? We have seen that it usually seems useful when somebody
gives you a dataset to work out three things to summarize it: the mean, the SD, and the histogram. Maybe that will help here, too.

"Wait," you will be thinking by now, if this whole thing is starting to seem farfetched to you (as well it may), "How can you figure out the mean and SD of an infinite number of values of a variable you've only imagined, much less the histogram?" It may sound impossible, but in the end it just turned out to be really hard: It took people about 150 years, from roughly 1650 to 1800, to answer these questions in full generality. The main actors in the drama were [xxx, ..., and xxx (for more of the history see Stigler, 19xx)]. I'll take the three items on the list—mean, SD, and histogram—in that order.

The Expected Value of a Sum. Let's talk about the mean of the imaginary dataset first, which I'm going to call the long-run average or expected value of the sum $S$—symbolically, $E(S)$, or $E_{\text{IID}}(S)$ if we're beingnotationally careful to remind ourselves of the sampling mechanism. To begin with, what ingredients should go into calculating this quantity? Well, remember what $S$ is supposed to stand for—it's your net gain after 100 gambles. The long-run average or expected value of $S$ should certainly depend on how much money you expect to win or lose on each gamble, which we called $\mu$ awhile ago and which in this case came out $-0.25$. It should also depend on how many gambles go into the sum, which we have called $n$—after all, if you expect to lose money each time, the more times you gamble the more you should expect to lose. So I guess the expected value of $S$ should involve $n$ and $\mu$, but how? Here's a way to put it that makes the answer clear:

Q: If I expect to lose a quarter each time I do something and I do it 100 times, how much money do I expect to lose overall?

A: A hundred quarters.

In other words,

The expected value of the sum of some IID draws from a population is the number of draws times the population mean. Symbolically,

\[ E(S) = E_{\text{IID}}(S) = n\mu . \]  

(9.6)
Here with \( n = 100 \) and \( \mu = -0.25 \) the expected value works out to \(-$25\), which answers part of the third question in the case study: After 100 $1 plays you expect to be behind by about $25. This looks like a good game for the casino.

**Math Interlude.** If you get interested in the random-variable end of this business you will want to know more about how to work algebraically with expected values. Here are two basic rules for doing so, together with a few words meant to make them plausible. In what follows the symbol \( c \) stands for a constant (something that does not vary randomly from one repetition to the next of whatever it is you’re studying) and upper-case symbols like \( Y \) are random variables (things that do vary randomly from one repetition to the next).

- \( E(cY) = cE(Y) \) (if you multiply something that is varying randomly around 5 (say) by 2 (say), the product should vary randomly around 10). In words, _multiplying a random quantity by a constant multiplies its expected value by the same constant._

- \( E(Y_1 + Y_2) = E(Y_1) + E(Y_2) \) (if one thing is varying randomly around 3 (say) and another around 4 (say), the sum should vary randomly around 7). This rule applies to adding more than just two random quantities together; in words you could say that _the expected value of the sum is the sum of the expected values_, whenever the number of terms going into the sum is finite.

These rules exactly parallel those back in Chapter 3 for working with means of data values. It’s easy to use the second rule to prove the expected value formula (9.6) above: with \( S = Y_1 + \ldots + Y_n \), in which each draw has mean \( \mu \),

\[
E(S) = E(Y_1 + \ldots + Y_n) = E(Y_1) + \ldots + E(Y_n) = \mu + \ldots + \mu = n\mu .
\] (9.7)

**The Standard Error of a Sum.** Well, after 100 plays you expect to be behind by $25, but give or take how much? Maybe the give-or-take is big enough that you still have a pretty good chance of coming out
ahead. This is where the SD of the imaginary dataset comes in. It has a special name—it’s called the **standard error** of the sum $S$, written $SE(S)$ or $SE_{IID}(S)$ if we’re being explicit about how the draws were chosen. The SE is harder to pin down than the expected value—we can use intuition to figure out what ingredients should go into it, and in roughly what way, but the precise formula requires some math to derive. Here we’ll just go through the intuition and I’ll say what the answer is, and you can read through the math interlude below for more information about where the formula comes from.

The issue is how much the sums in the imaginary dataset should vary around their long-run mean $-$9.25. In figuring this out it helps to remember that each value of $S$ is the sum of 100 IID draws from the population, all of which are fluctuating around the population mean $\mu = -$0.25 by an amount given by the population SD $\sigma$, which here came out $1.30$. It makes pretty good intuitive sense that the variability of the sum should depend on the variability of the draws going into the sum, so I guess $SE(S)$ should involve $\sigma$. Moreover the standard error should involve the population SD in such a way that as $\sigma$ goes up, so does the SE, because as the individual draws become more variable, their sum should, too. We will see as we go along that standard error formulas often take the form of fractions, with a numerator and a denominator, and that the ingredients in the formulas tend to appear either in the numerator or the denominator but not in both. I have just argued intuitively that $\sigma$ should appear in the numerator of $SE(S)$. This also makes sense by thinking about the *units* in which $\sigma$ and the SE are measured: in Keno $\sigma$ is in dollars, the SE (which is after all just the SD of the sums in the imaginary dataset) should also be in dollars, and it would be a mistake to put $\sigma$ in the denominator because then the units would come out wrong.

It turns out that there is just one other ingredient in the formula for $SE(S)$: the number $n$ of draws going into each sum. You can see that $n$ is relevant to the SE by thinking over the following three facts:

- Noticing that the sums in the imaginary dataset exhibit variability around their mean is equivalent to saying that we are uncertain about the precise value that any one of the sums will take on,

- This uncertainty arises because we are in turn uncertain about
the exact value each of the draws going into a given sum will take, and

- It stands to reason that our uncertainty about a sum should depend in part on how many uncertain things are being added together to yield the sum.

Moreover, this same argument shows that $n$ should appear in the numerator of the standard error formula along with $\sigma$, because the more things you add together, each of which is uncertain, the more uncertainty you should have about the sum.

That’s about as far as intuition can take us; I now have to just say the answer, which comes out like this:

The standard error of the sum of some IID draws from a population is the population SD times the square root of the number of the draws. Symbolically,

\[
SE(S) = SE_{\text{IID}}(S) = \sigma \sqrt{n}.
\] (9.8)

Here with $\sigma = 1.30$ and $n = 100$ the SE works out to $1.30 \cdot \sqrt{100} = 13$. Putting this together with the expected value answers another part of the third question in the case study: the interpretation is that after 100 plays you expect to be behind by about $25$, give or take about $13$. In terms of the sums in the imaginary dataset this means that you would not be surprised to see a sum like $-31$ or $-10$ but you would be pretty surprised to see a value like $+20$, because $+20$ is $\frac{+20-(-25)}{13} = 3.5$ SDs above average and that sort of thing doesn’t happen very often. It’s beginning to sound like our chance of coming out ahead with single-number betting in Keno isn’t too good.

Math Interlude. You will also want to know a few rules for algebraically manipulating standard errors if you get into the random variables story more deeply. It turns out that it is easier to express what’s going on not in terms of the SE of a random quantity but in terms of the square of the SE, which—since a standard error is just a kind of standard deviation—people call the variance of the random quantity, in parallel with the terminology in Chapter 3 for measures of spread for data values.
Let’s use the notation \( V(Y) \) to stand for the variance of a random variable \( Y \), so that symbolically the definition in the last sentence becomes \( V(Y) = [SE(Y)]^2 \) and \( SE(Y) = \sqrt{V(Y)} \). As with expected values there are two basic rules to learn about how variances work, and each rule can then be translated back into a statement about standard errors, by taking the square root of both sides of the variance formula.

- \( V(cY) = c^2V(Y) \), so that \( SE(cY) = |c|SE(Y) \) (if you multiply something that is varying randomly around its mean with a give-or-take of 3 (say) by 2 (say), the result should vary around its mean with a give-or-take of 6). In words, multiplying a random quantity by a constant multiplies its variability by the absolute value of the constant.

- If \( Y_1 \) and \( Y_2 \) are independent, \( V(Y_1 + Y_2) = V(Y_1) + V(Y_2) \), so that in this case

\[
SE(Y_1 + Y_2) = \sqrt{V(Y_1 + Y_2)} = \sqrt{V(Y_1) + V(Y_2)} = \sqrt{[SE(Y_1)]^2 + [SE(Y_2)]^2}.
\]

(9.9)

In words, When independent random quantities are added together, the variance of the sum is the sum of the variances, and the standard error of the sum follows a Pythagorean law: the SE is like the hypotenuse of a right triangle whose sides are the SE's of the terms going into the sum. We will return to this fact later in Chapter xx, where it is crucial to analyzing data gathered with two independent samples from different populations. Notice that it is on the variance scale that variability is additive, not the standard error scale. This is the only reason the term “variance” was invented in the first place—after all, as a measure of variability it has the wrong units (if \( Y \) is in dollars then \( V(Y) \) is in dollars\(^2\), which are pretty hard to spend). In practice people do math about the uncertainty expressed by random variables on the variance scale and then interpret the results by taking the square root, to get things back onto the standard error scale where they belong.
If you spot me without proof the formula for the variance of a sum, we can now see immediately where the formula \( SE(S) = \sigma \sqrt{n} \) comes from: since \( S = Y_1 + \ldots + Y_n \) and the individual draws \( Y_i \) are IID, meaning that the \( Y_i \) are independent and all have the same standard error (SD) \( \sigma \) and hence variance \( \sigma^2 \),

\[
V(S) = V(Y_1 + \ldots + Y_n) = V(Y_1) + \ldots + V(Y_n) = \sigma^2 + \ldots + \sigma^2 = n\sigma^2 ,
\]

and you get the result we want, \( SE(S) = \sigma \sqrt{n} \), by taking the square root of both sides of this equation.

9.3 The Central Limit Theorem

This is a long calculation, and we're almost to the end of it. We're still wondering what \( P(S > 0) \) is, which is equivalent to wondering what percentage of the sums in the imaginary dataset are positive. If we knew what the histogram of the sums in that dataset looked like (drawn on the density scale for convenience), we could answer this question by working out the area under the histogram to the right of \( 0 \), in much the same way that Quetelet answered questions about the heights of his Scottish soldiers back in Chapter 4. So, what does the histogram of the sum of a bunch of IID draws from a population look like?

Of the three summaries of the imaginary dataset we have considered (mean, SD, and histogram), this was the hardest for people to figure out. It turns out that the answer depends on two things—the sample size \( n \) and the shape of the population histogram—which at first glance makes this whole approach sound impractical again, since there's an infinite variety of possible shapes the population histogram could take on and an infinity of possible values for \( n \). But in XXX, and people building on his work, were able to prove the following remarkable result, which has come to be called the Central Limit Theorem, or CLT for short:

**Central Limit Theorem:** Pretty much no matter what the population histogram looks like, as long as the number \( n \) of IID draws going into a sum is big enough, the histogram of the sum \( S \) will follow the normal curve pretty well.
This is one of the two most important theorems in the book, along with Bayes' Theorem—we will see that it makes feasible many probability calculations that would otherwise have been quite hard to carry out. Let’s use the CLT to finally compute the chance of coming out ahead with 100 single-number $\$1$ gambles at Keno, and then I’ll have more to say about when the theorem applies in practice and when it doesn’t.

The picture in the lower right corner of Figure 1 finishes the calculation: according to the CLT, the histogram of the sum follows the normal curve with a mean (expected value) of $-\$25$ and an SD (standard error) of $\$13$ pretty well, so $Z = \frac{\$0 - (-\$25)}{\$13} = 1.9$ and the area to the right of 1.9 under the standard normal curve from Table A.1 is about 2.7%, or about 1 chance in 37. If you play single-number Keno once for $\$1$, your chance of coming out ahead is only 1 in 4; if you persevere and make 100 such plays, your chance of coming away a winner drops to only 1 in 37. Keno is a terrible game for the gamblers and a great game for the casino.

Math Interlude. The way I stated the Central Limit Theorem a minute ago makes it sound pretty vague, but it has a precise statement as well:

If $Y_1, \ldots, Y_n$ are IID draws from a population with mean $\mu$ and finite SD $\sigma$, $S = Y_1 + \ldots + Y_n$, and $\Phi(z)$ is the function that keeps track of the area under the standard normal curve to the left of $z$, then for all $z$

$$
\lim_{n \to \infty} P\left(\frac{S - n\mu}{\sigma\sqrt{n}} \leq z\right) = \Phi(z). \quad (9.11)
$$

This theorem has hundreds of variations—as you read this page, I'm sure that someone somewhere is proving another form of it—and (unlike Bayes' Theorem) the plain-vanilla version I just stated requires some fairly high-powered math to demonstrate (most people taking statistics courses don't see a complete proof of it until they are graduate students in the subject). It's a pleasant fact that we don't need to know how to prove it to make good use of it in the rest of the book.

Figure 2a shows the actual histogram of the sum of 100 IID draws from the population in Figure 1, with the normal curve specified by
the Central Limit Theorem superimposed on it. The real histogram has gaps in it, because you can't get all possible numbers from $-\infty$ to $+\infty$ by adding together 100 -$1$'s and $+$2's; in fact the smallest possible value for your net gain is $-100$ (if you lost every time), the second-smallest possible value is $-97$ (if you won only once), and so on up to the biggest possible value of $+200$ (if you were so lucky as to win every time), with the possible net gains occurring every $3$ along the way from $-100$ to $+200$ ($3$ comes from the fact that $-1$ and $+2$ are $3$ apart). The CLT does not appear to provide a very good normal approximation here, but this is visually misleading since the normal curve is continuous and the real histogram is discrete. When I correct for this in Figure 2b, by replotting the real histogram with bars that are $3$ wide and adjusting the normal curve accordingly, you can see that in fact the normal approximation is quite good—for example, the exact value of $P(S > 0)$ in this case is $2.75\%$, which is close to the approximate value we got above from the CLT ($2.71\%$).

Evidently for this population histogram and this value of $n$ the CLT has worked pretty well. But the theorem (especially as I stated it above) is unclear on an important point—how big does $n$ have to be to get a good normal approximation for the histogram of a sum?

When Does the Central Limit Theorem Apply? This question has no easy general answer. There's a hint from the word "limit" in the name of the theorem—mathematically the CLT guarantees a perfect normal curve only in the limit as the number $n$ of draws going into each sum $S$ becomes infinitely large. Fortunately perfection is not needed, and—as the single-number Keno example above in Figure 2 demonstrates—values of $n$ considerably smaller than $+\infty$ can yield really good normal approximations. One thing should be clear, though, from the idea of taking a limit as $n$ goes to infinity: The normal approximation for the histogram of the sum improves as $n$ grows. If you already have a good normal approximation with $n = 100$, as in Figure 2, it would be even better for $n = 200$.

Note that what counts in getting a good normal approximation for the histogram of the sum is not the population size
$N$, or the number $M$ of hypothetical sums in the imaginary dataset; it's the number $n$ of draws going into each sum.

The best way to gain additional insight into when the CLT applies is by looking at a few more examples. Here are two.

- **What if the population histogram is normal to begin with?** Suppose the population has many values in it, and a histogram of those values follows the normal curve pretty well, as in Figure 3. Then the following math fact comes in handy:

  **Math Fact:** For any $n$, the histogram of the sum of $n$ IID draws from a normal population is itself normal.

  This means that when the population histogram follows the normal curve pretty well to begin with, the histogram of the sum will already follow the normal curve well even with only $n = 1$! This makes sense if you think about it: With only one value $Y_1$ going into the “sum,” the sum just is that value, so the histogram of the sum is just the histogram you would get by repeatedly taking one draw at random from the population and then replacing it—but that’s just the sample histogram, which we already know is supposed to look like the population histogram, which in this case is normal. Evidently

  The closer the population histogram is to normality to begin with, the smaller $n$ needs to be to get a good normal approximation for the histogram of the sum.

- **What if the population histogram is quite skewed?** Figure 4a gives the histogram of hospital expenditures for Medicare patients in 19xx with a heart attack. You can see it has a quite long right-hand tail—the typical Medicare patient with a heart attack ran up a hospital bill of about $xxxx$, but some people had much higher expenses than that. An average hospital will have about 50 such patients per year. If we think of Figure 4a as a population histogram and imagine taking an IID sample of $n = 50$ of these patients, the sum of those 50 draws would keep track for us of the total amount of money paid out by the government per
year to a typical hospital to treat heart attacks in elderly people, which would be a dollar figure of some policy interest. Of course, this total would vary from hospital to hospital according to some histogram. What would this histogram look like?

Figure 4b shows the exact histogram, together with the normal approximation offered by the CLT. The normal curve doesn’t fit too well in this case—with only 50 draws going into each sum, the histogram of the sum is still pretty skewed. Figure 4c shows the corresponding histograms for $n = 200$, and you can see that the normal approximation is much better. The moral is that

*If the population histogram is sharply skewed, $n$ needs to be quite a bit bigger to get a good normal approximation for the histogram of the sum.*

To summarize all of this,

$n = 50$ or 100 should be enough to give a workable normal approximation for the histogram of the sum of a bunch of IID draws, unless the population histogram is strongly skewed to begin with, in which case several hundred draws should be enough. Some populations are close enough to normality to begin with that $n = 5$ or 10 draws are already enough.

All general rules of this type can be violated—in other words, you can invent populations for which this rule makes you think you have a better normal approximation than you actually do—but we will see as we go along in the book that building probability models to learn about real-world phenomena is a process that involves several approximations to reality, and when the modeling process fails it rarely does so because the CLT has let you down. I will return to this point in later chapters.

**Comparing Single-Number and Double-Number Gambling at Keno.** Let’s wrap up the chapter by working out the chance of coming out ahead for double-number gambling at Keno—this will review
all the steps in building a probability model, and there is an interesting punchline about how to decide which gamble is better. Figure 5 summarizes the results of the modeling.

First the population. Recall ages ago at the beginning of the chapter I said that the payoff for double-number wagering at Keno was 11 to 1. If we're going to imagine making $1 bets, as we did with the single-number gambling strategy, this means that the population should have -$1's and +$11's in it, but how many of each? You can verify using logic similar to that in Section 1 that, just as the chance of winning any single number bet was \( \frac{20}{60} \), the chance of winning a double-number bet is \( \frac{36}{60} = \frac{3.6}{6} \), or almost exactly 6% (see Problem xx). If we're being careful in building the population dataset we should put 380 +$11's and (6320 - 380) = 5940 -$1's in it, which is what I did in Figure 5, but you could get a good approximation to this by using 6 +$11's and 94 -$1's if that seems simpler to you.

Next comes the population mean \( \mu \) and SD \( \sigma \):

\[
\mu = \frac{380(+$11) + 5940(-$1)}{6320} = \frac{-$1760}{6320} = -$0.28,
\]

\[
\sigma = \sqrt{\frac{380[+11 - (-$0.28)]^2 + 5940[-1 - (-$0.28)]^2}{6320}} = $2.85.
\]

Double-number gambling is a bit worse than single-number wagering in how much money you expect to lose each time (28¢ versus 25¢), but it's quite a bit more variable in its outcomes (the SD is more than twice as big). What do you think these two facts imply about whether your chance of coming out ahead is bigger with the double numbers? ♠♠

The sample part of the model diagram is identical to that with single-number gambling—\( n = 100 \) IID draws, with attention focusing on the sum—except that the sample histogram should now resemble the population histogram on the left side of Figure 5, which is considerably more skewed than its counterpart back in Figure 1. The imaginary dataset stands for the same thing as it did earlier, with each entry representing a possible sum of 100 draws, but its long-run mean (the expected value of the sum) and SD (the standard error) are different:

\[
\text{expected value} = E(S) = n\mu = 100(-$0.28) = -$28,
\]

\[
\text{standard error} = SE(S) = \sigma\sqrt{n} = 10 \cdot $2.85 = $28.50. \tag{9.12}
\]
I expect to lose about $28 with double-number gambling, which is not much different than with single-number betting ($25), but the give or take around the expected value is quite a bit bigger ($28.50 versus $13). This means, interestingly, that my probability of coming out ahead, $P(S > 0)$, is quite a bit bigger than it was with the single numbers, as follows: (1) the Central Limit Theorem should again give us a pretty good normal approximation to the histogram of the sum; (2) "coming out ahead" with double-number gambling works out to $\frac{$30-(-$28)}{$28.50} = 0.98$ in standard units; and (3) $P(S > 0)$ therefore comes out to about 16%. (The CLT doesn't give quite as good an approximation here—the exact answer is about 14.8%—but it's good enough for our purposes.) A coming-out-ahead chance of 15% or so is more than five times larger than the corresponding value for single-number wagering (2.7%), so I guess double-number betting at Keno is better. Or is it?

Figure 6 shows the normal approximations to the two histograms for your likely net gain after 100 plays with the two strategies, plotted on the same dollar scale. The double-number approach does indeed put more area under the curve to the right of $0$, because double-number wagering is more variable (more uncertain) than single-number betting, but the something-for-nothing bell should be going off in your head if I tried to tell you that this benefit came without any cost. The figure shows what the cost is: the additional variability created by double-number betting means that your chance of losing a lot of money is also a lot bigger than it was with the single numbers. For instance, with single-number betting your chance of losing $50 or more is the same as your chance of coming out ahead—2.7%—because $-50$ is just as far to the left of the expected value ($-25$) as $0$ is to the right of it, but with double-number wagering this chance is well over 20%.

So which gamble is better? There is no single "correct" answer for everybody—it depends on how conservative you are about taking risks. People studying decision theory would call you risk-averse or risk-seeking according to your preference in situations like this (see the discussion in Chapter xx about utility functions). Which gamble do you like better? Or is the punchline in this case that you should keep your money in your pocket?
9.4 Chapter Summary

1. Probability models are useful for studying random processes like the amount of money you're likely to win or lose at games of chance. Such models have three parts: the population, the sample, and the imaginary dataset. Each can be thought of as a dataset, with mean, SD, and histogram. The population represents what you might get each time you repeat the process you're studying; the sample helps to simulate a number of such repetitions, and to focus attention on a summary of the repetitions like their sum; and the imaginary dataset helps to visualize what you would get if you took repeated samples and wrote down the summary value each time. See Figures 1 and 5 for examples of probability modeling diagrams that bring all these ingredients together.

2. To specify the population, you identify the possible values the random process you're studying could produce on any given repetition, together with the probabilities associated with each of those values. In gambling models, for instance, the possible values are the dollar amounts that could change hands each time the gamble is repeated, and you can often use the rules in Chapter 7 to figure out the probabilities of each of these outcomes occurring. The population mean \( \mu \) and SD \( \sigma \) are computed in the same way we did back in Chapter 3 with any other lists of numbers, except that when working out the population SD you divide by the number \( N \) of values in the population rather than \( (N - 1) \).

3. The sample represents one possible outcome of a bunch of repetitions of the random process you're studying. In the Keno example of Section 1, for instance, we were wondering what would happen if you played Keno 100 times, so the population modeled the individual plays and the sample collected \( n = 100 \) such gambles together. It's important when describing the sample to be specific about the sampling mechanism—the relationship between the sample and the population, of which IID (drawing the sample at random with replacement, as in gambling) and SRS (at random without replacement) are the two simplest cases. The other important thing to specify about the sample is what summary of it is of particular interest in the problem at hand. In the Keno models of this chapter, for instance, the individual draws represented the amounts of money we won or lost on each individual
Keno game, and the sum of the draws in the sample stood for our net gain after all 100 gambles.

4. The imaginary dataset helps you visualize what would happen if you took repeated samples and calculated the relevant sample summary each time. With Keno, for example, you imagine making 100 IID draws and calculating their sum S, making 100 more IID draws and working out their sum, and so on, writing down the sums you would get in a dataset. Calculating the chance of coming out ahead after 100 gambles—symbolically, \( P(S > 0) \)—then amounts to working out the percentage of the sums in the imaginary dataset that are positive, or equivalently computing the area under the histogram of the imaginary dataset to the right of 0.

5. This area is hard to work out exactly, but—when the summary of interest is the sum of some IID draws from the population, as it was with Keno—the probability we want may often be approximated by computing the long-run mean and SD of the imaginary dataset and working out the relevant area under the normal curve with that mean and SD.

6. The long-run mean of the sums in the imaginary dataset is called the expected value of the sum S, written \( E(S) \). You can calculate it with the following rule:

\[
\text{The expected value of the sum of some IID draws from a population is the number of draws times the population mean. Symbolically, } E(S) = n \mu.
\]

7. The long-run SD of the sums in the imaginary dataset is called the standard error of the sum S, written \( SE(S) \). The rule for calculating it is as follows:

\[
\text{The standard error of the sum of some IID draws from a population is the population SD times the square root of the number of draws. Symbolically, } SE(S) = \sigma \sqrt{n}.
\]

8. As long as the number \( n \) of IID draws going into a sum is big enough, the histogram of the sum S will follow the normal curve pretty well. This result is called the Central Limit Theorem, or CLT for short. The closeness of the normal curve to the actual histogram of the sum
depends on two things: the sample size $n$ and the population histogram. The closer the population histogram is to normality to begin with, the smaller $n$ needs to be to get a good normal approximation. In general, $n = 50$ or 100 should be enough unless the population histogram is quite skewed, in which case several hundred draws going into each sum may be needed to get a good normal curve. Some populations look normal enough to begin with that already with only $n = 5$ or 10 you have a nice normal approximation.

9. (Math) The mathematical objects that stand for things like the process of drawing at random from a population are called random variables. In the Keno example, for instance, $Y_2$ stood for the process of making the second draw from the population in Figure 1, and $S$ stood for the process of making all 100 draws and computing their sum. $Y_2$ and $S$ are both random variables, because each time you make the second draw or add together all 100 draws you will get a different result at random. Random variables have means and SDs just like datasets, but the names are different: the mean of a random variable is called its expected value, and the SD is called its standard error. These terms coincide with the idea, expressed more informally in this chapter, of the long-run mean and SD of the imaginary dataset. Expected values and standard errors of random variables obey the following simple rules, in which $c$ stands for a constant and $V(Y)$ is the square of the standard error of $Y$ (the variance of $Y$):

- $E(cY) = cE(Y)$;
- $E(Y_1 + Y_2) = E(Y_1) + E(Y_2)$;
- $V(cY) = c^2V(Y)$, so that $SE(cY) = |c|SE(Y)$; and
- If $Y_1$ and $Y_2$ are independent, $V(Y_1 + Y_2) = V(Y_1) + V(Y_2)$, so that $SE(Y_1 + Y_2) = \sqrt{[SE(Y_1)]^2 + [SE(Y_2)]^2}$.

9.5 Problems

1. (gambling) What should the payoff have been in single-number wagering at Keno so that the gamble is fair? What about with double-number betting? If both gambles were fair, show that neither would have an advantage over the other with respect to the probability of coming out ahead. What other criteria would you use, and which of
the two fair gambles seems better to you when evaluated with your criteria? Explain briefly.

2. (engineering) Some traffic engineers working for the city of Los Angeles are trying to synchronize the lights on a long stretch of Santa Monica Boulevard so that someone driving along at a constant speed like 30 miles an hour would hit them all green (you can tell this problem is hypothetical—they would never do anything that sensible), so they set up a traffic survey in which they record the times at which cars pass various points along the street at various times of the day and night on weekdays and weekends. There are two main ways to study data of this type—you can count the number of cars going by in fixed-length intervals of time, or you can keep track of the times between cars. The point of this problem is that even if you choose just one of these methods you can answer questions phrased in terms of the other, so in that sense the two approaches are equivalent.

These engineers decide to focus on the times between cars, and examination of the data for one spot along the street in one direction shows that during one stable time period, on weekday afternoons from 1 to 3:30 pm, there is no systematic trend or pattern to the interarrival times and they average 5.5 seconds with an SD of 4 seconds.

One of the standard volume-of-traffic criteria used in studies like this involves the number of cars passing by in a 10-minute period: if at least 100 cars often go by in that amount of time, the spot is characterized as “busy” and treated differently in the subsequent analysis. Suppose that a car passes by this spot one day at some moment in this weekday afternoon period, and one of the engineers starts her stopwatch at that moment and begins counting cars. How likely is it that at least 100 cars go by in the next 10 minutes? Would you classify this as a busy spot? Explain briefly. (Hint: This problem is not straightforward. Try building a probability model based on the times between cars, and relate the event that at least 100 cars go by in 10 minutes to this model.)

3. (gambling) Show, using reasoning similar to that in Section 1, that the chance of winning any given double-number bet in Keno is \( \frac{29 \times 80}{80\text{-}79} = \frac{380}{820} \), or just about exactly 6%. How does this generalize to three numbers, or more? Explain briefly.
4. (probability modeling) From what you have seen in this chapter, why do people concentrate on the mean and SD as summaries of the population, sample, and imaginary datasets, rather than other measures of center and spread from Chapter 3 like the median and interquartile range? Explain briefly. (*Hint:* The Medicare example in Figure 4a is helpful (what is the quantity of greatest real-world interest in that example?), and so is the CLT.)

5. (gambler's ruin)

6. (lottery)

7. (roulette)

8. (the elevator problem?; more)
Figure 2: A probability model for single-number $2$ betting at the.

**Population:**
- All possible net gains

**Sample:**
- The observed outcome

**Sample Data Set:**
- Possible $S$ values

$$\begin{bmatrix}
-81 \\
-81 \\
-91 \\
+62
\end{bmatrix}$$

$$\sum S = -10$$ (say)

$$\bar{S} = \frac{-10}{100} = -0.1$$

$$\text{mean } \mu = -0.25$$

$$\text{SD } \sigma = 1.30$$

$$\frac{3}{4} \text{ pop. hist.}$$

$$-1 \, \, 0 \, \, 1 \, \, 2$$

$$\text{A table of areas under the normal curve (from the book of any probability or statistics book) says that the area of interest here is about 2.1%}.$$