17
Probability models

Describing randomness

The basic idea of probability is that even random outcomes exhibit structure and obey certain rules. In this chapter, we'll learn to use these rules to build probability models, which employ the language of probability theory to provide mathematical descriptions of random phenomena. Probability models can be used to answer interesting questions about uncertain real-world systems. For example:

- American Airlines oversells a flight from Dallas to New York, issuing 140 tickets for 134 seats, because they expect at least 6 no-shows (i.e. passengers who bought a ticket but fail to show up for the flight). How likely is it that the airline will have to bump someone to the next flight?

- Arsenal scores 1.6 goals per game; Manchester United scores 1.3 goals per game. How likely it is that Arsenal beats Man U when they play each other?

- Since 1900, stocks have returned about 6.5% per year on average, net of inflation, but with a lot of variability around this mean. How does this variability affect the likely growth of your investment portfolio? How likely it is that you won't meet your retirement goals with your current investment strategy?

Building a probability model involves two simple steps.

First, you identify the *random variables* of interest in your system. A random variable is just a **numerical summary of an uncertain outcome.** In the airline example, we could have any possible combination of passengers fail to show up (seat 2C, 14G, etc). But at the end of the day, if we want to know whether any passengers are likely to get bumped to the next flight, all we care about is *how* *many* ticketed passengers are no-shows, not their specific identities or seat numbers. So that's our numerical summary, i.e. our random variable: X = the number of no-shows. Or in the soccer game between Arsenal and Man U, there are two obvious numerical summaries: X_1 = the number of goals scored by Arsenal, and X_2 = the number of goals scored by Man U.

Second, you provide a rule for calculating probabilities associated with each possible value of the random variable. This rule is called a *probability distribution*. In the airline example, this distribution might be described using a simple lookup table based on historical data, e.g. 1% of all flights have 1 no-show, 1.2% have 2 no-shows, 1.7% have 3 no-shows, and so forth. In building a probability model, this second step is usually where the action is, and it's what we'll discuss extensively in this chapter.

There are two common types of random variables, corresponding to two common types of outcomes.

Discrete: the possible outcomes are whole numbers (1, 2, 3, etc.). Both the number of airline no-shows and the score of a soccer game are discrete random variables: you can't have 2.4 no-shows or 3.7 goals.

Continuous: the random variable could be anything within a continuous range of numbers, like the price of Apple stock tomorrow, or the volume of subsurface oil reservoir.

An example. Here's a simple example that will help you practice your understanding of these concepts. Imagine that you've just pulled up to your new house after a long cross-country drive, only to discover that the movers have buggered off and left all your furniture and boxes sitting in the front yard. What a mess! (This actually happened to a friend of mine.) You decide to ask your new neighbors for some help getting your stuff indoors. Assuming your neighbors are the kindly type, how many pairs of hands might come to your aid? Let's use the letter X to denote the (unknown) size of the household next door. The table at right shows a probability distribution for X, taken from U.S. census data in 2015; you might find this easier to visualize using the barplot in Figure 17.1.

Size of household, x	Probability, $P(X = x)$
1	0.280
2	0.336
3	0.155
4	0.132
5	0.060
6	0.023
7	0.011
8	0.003

Table 17.1: Probability distribution for household size in the U.S. in 2015. There is a vanishingly small probability for a household of size 9 or higher, which is just rounded off to zero here.

This probability distribution provides a complete representation of your uncertainty in this situation. It has all the key features of any probability distribution:

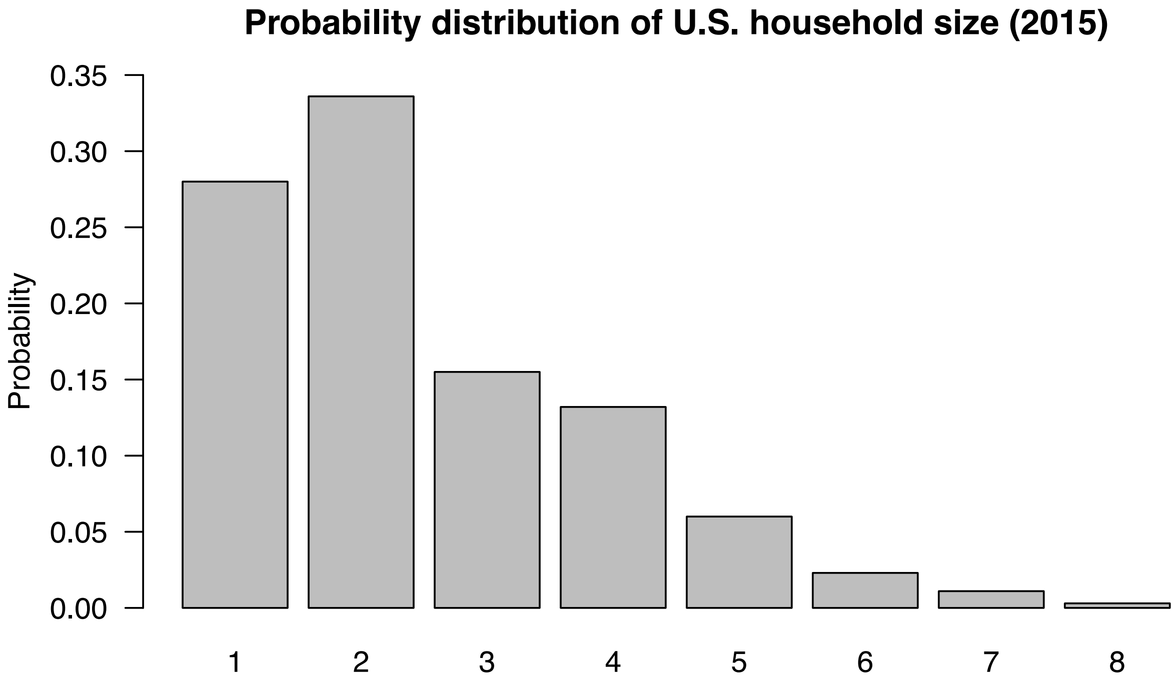


Figure 17.1: Probability distribution for the size of a random U.S. household in 2015. The elements of the sample space (the numbers $x = 1$ through $x = 8$) are shown along the horizontal axis. The probabilities $P(X = x)$ are shown on the vertical axis.

1. There is a random variable, or a numerical summary of an uncertain situation—here, the size of the household next door (X).

2. There is a set of possible outcomes for the random variable—here, the numbers 1 through 8.

3. Finally, there are probabilities for each possible outcome—here provided via a simple look-up table. Notice that the table uses big X to denote the random variable itself, and little x to denote the possible outcomes.

Most probability distributions won't be this simple, but they will all require specifying these same basic elements.

Expected value: the mathematical definition

When you knock on your neighbors' door in the hopes of getting some help with your moving fiasco, how many people should you "expect" to be living there?

The *expected value* of a probability distribution for a numerical random variable is just an average of the possible outcomes—but a *weighted* average, rather than an ordinary average. This is a crucial distinction. If you take the 8 possible outcomes in Figure 17.1 and form their ordinary average, you get

$$\text{Ordinary average} = \frac{1}{8} \cdot 1 + \frac{1}{8} \cdot 2 + \cdots + \frac{1}{8} \cdot 7 + \frac{1}{8} \cdot 8 = 4.5\,.$$

Here, the weight on each possible outcome is $1/8 = 0.125$, since there are 8 numbers. This is *not* the expected value; it give each possible outcome an equal weight, ignoring the fact that these numbers have different probabilities.

To calculate an expected value, we instead form an average using *unequal* weights, given by the probabilities of each outcome:

$$\text{Expected value} = (0.280) \cdot 1 + (0.336) \cdot 2 + \cdots + (0.011) \cdot 7 + (0.003) \cdot 8 \approx 2.5\,.$$

The more likely numbers (e.g. 1 and 2) get higher weights than $1/8$, while the unlikely numbers (e.g. 7 and 8) get lower weights.

This example conveys something important about expected values. Even if the world is black and white, an expected value is often grey. For example, the expected American household size is 2.5 people, a baseball player expects to get 0.25 hits per at bat, and so forth.

As a general rule, suppose that the possible outcomes for a random variable X are the numbers $x_1, \ldots, x_N$. The formal definition for the expected value of X is

$$E(X) = \sum_{i=1}^{N} P(X = x_i) \cdot x_i\,. \tag{17.1}$$

This measures the "center" or mean of the probability distribution. Later, we'll learn how this more formal definition of expected value can be reconciled with the NP rule—that is, with our previous understanding of expected value as a risk/frequency calculation.

A related concept is the *variance*, which measures the dispersion or spread of a probability distribution. It is the expected (squared) deviation from the mean, or

$$\text{var}(X) = \text{E}\big(\{X - \text{E}(X)\}^2\big)\,.$$

The standard deviation of a probability distribution is $\sigma = \text{sd}(X) = \sqrt{\text{var}(X)}$. The standard deviation is more interpretable than the variance, because it has the same units (dollars, miles, etc.) as the random variable itself.

Parametric models for discrete outcomes

Of the steps required to build a probability model, the requirement that we provide a rule that can be used to calculate probabilities for each possible outcome is usually the hardest one. In fact, for most scenarios, if we had to build such a rule from scratch, we'd be in for an awful lot of careful, tedious work. Imagine trying to list, one by one, the probabilities for all possible outcomes of a soccer game, or all possible outcomes for the performance of a portfolio containing a mix of stocks and bonds over 40 years.

Thus instead of building probability distributions from scratch, we will rely on a simplification called a *parametric probability model.* A parametric probability model involves a probability distribution that can be completely described using a relatively small set of numbers, far smaller than the sample space itself. These numbers are called the parameters of the distribution. There are lots of commonly used parametric models—you might have heard of the normal, binomial, Poisson, and so forth—that have been invented for specific purposes. A large part of getting better at probability modeling is to learn about these existing parametric models, and to gain an appreciation for the typical kinds of real-world problems where each one is appropriate.

Recall our distinction earlier between *discrete* and *continuous* random variables. A discrete random variable means that you can count the possible outcome on your fingers and toes.[1] Examples here include the number of no-shows on a flight, the number of goals scored by Man U in a soccer game, or the number of gamma rays emitted by a gram of radioactive uranium over the next second. Continuous random variables, on the other hand, can take on any value within a given range, like the price of a stock or the speed of a tennis player's serve.

We'll start with the case of discrete random variables. Suppose that we have a random variable X whose possible outcomes are x_1, x_2, and so forth. You'll recall that, to specify a probability model, we must provide a rule that can be used to calculate $P(X = x_k)$ for each possible outcome. When building parametric probability models, this rule takes the form of a *probability mass function*, or PMF:

$$P(X = x_k) = f(x_k \mid \theta) \, .$$

In words, this equation says that the probability that X takes on the value x_k is a function of x_k. The probability mass function

[1] Repeating when you get beyond 20 :-)

depends a number (or set of numbers) θ, called the parameter(s) of the model.

To specify a parametric model for a discrete random variable, we must both provide both the probability mass function f and the parameter θ. This is best illustrated by example. We'll consider two: the binomial and Poisson distributions.

The binomial distribution

One of the simplest parametric models in all of probability theory is called the binomial distribution, which generalizes the idea of flipping a coin many times and counting the number of heads that come up. The binomial distribution is a useful parametric model for any situation with the following properties:

(1) We observe N different random events, each of which can be either a "yes" or a "no."

(2) The probability of any individual event being "yes" is equal to P, a number between 0 and 1.

(3) Each event is independent of the others.

(4) The random variable X of interest is total number of "yes" events. Thus the sample space is the set $\{0, 1, \ldots, N-1, N\}$.

The meaning of "yes" events and "no" events will be context-dependent. For example, in the airline no-show example, we might say that a "yes" event corresponds to a single passenger failing to show up for his or her flight (which is probably not good for the passenger, but definitely a success in the eyes of an airline that's overbooked a flight). Another example: in the PREDIMED study of the Mediterranean diet, a "yes" event might correspond to single study participant experiencing a heart attack.

If a random variable X satisfies the above four criteria, then it follows a binomial distribution, and the PMF of X is

$$P(X = k) = f(k \mid N, P) = \binom{N}{k} P^k \, (1-P)^{N-k} \, , \tag{17.2}$$

where N and P are the parameters of the model. The notation $\binom{N}{k}$, which we read aloud as "N choose k," is shorthand for the following expression in terms of factorials:

$$\binom{N}{k} = \frac{N!}{k!(N-k)!} \, .$$

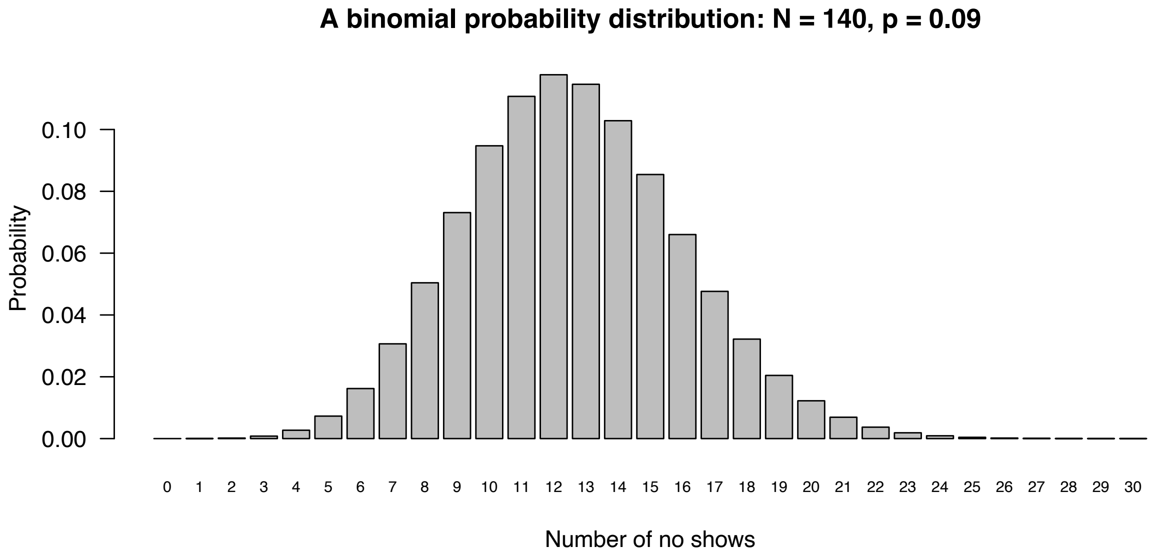


Figure 17.2: A barplot showing the probability distribution for the number of no-shows on an overbooked airline flight with 140 tickets sold, assuming a no-show rate of 9% and that individual no-shows are independent. The horizontal axis has been truncated at $k = 30$ because the probability of more than 30 no-shows is vanishly small under the binomial model.

This term, called a binomial coefficient, counts the number of possible ways there are to achieve k "yes" events out of N total events. (You'll see how this is derived in a bit.)

Example: airline no-shows Let's use the binomial distribution as a probability model for our earlier example on airline no-shows. The airline sold tickets 140 people, each of which will either show up to fly that day (a "yes" event) or not (a "no" event). Let's make two simplifying assumptions: (1) that each person decides to show up or not independently of the other people, and (2) that the probability of any individual person failing to show up for the flight is 9%.[2] These assumptions make it possible to apply the binomial distribution. Thus the distribution for X, the number of ticketed passengers who fail to show up for the flight, has PMF

$$P(X = k) = \binom{140}{k} \, (0.09)^k \, (1 - 0.09)^{140-k} \, .$$

This function of k, the number of no-shows, is plotted in Figure 17.2. The horizontal axis shows k; the vertical axis shows $P(X = k)$ under the binomial model with parameters $N = 140, p = 0.09$.

According to this model, the airline should expect to see around $E(X) = Np = 140 \cdot 0.09 = 12.6$ no shows, with a standard deviation of $\text{sd}(X) = \sqrt{140 \cdot 0.09 \cdot (1 - 0.09)} \approx 3.4$. But remember that

[2] This is the industry average, quoted in "Passenger-Based Predictive Modeling of Airline No-show Rates," by Lawrence, Hong, and Cherrier (SIGKDD 2003 August 24-27, 2003).

the question of interest is: what is the probability of fewer than 6 no-shows? If this happens, the airline will have to compensate the passengers they bump to the next flight. We can calculate this as

$$P(X < 6) = P(X = 0) + P(X = 1) + \cdots + P(X = 5) \approx 0.011\,,$$

i.e. by adding up the probabilities for 0 no-shows through 5 no-shows. The airline should anticipate a 1.1% chance that more people will show up than can fit on the plane.

The trade-offs of the binomial model. It's worth noting that real airlines use much more complicated models than we've just built here. These models might take into account, for example, the fact that passengers on a late connecting flight will fail to show up together non-independently, and that business travelers are more likely no-shows than families on a vacation.

The binomial model—like all parametric probability models—cannot incorporate these (very real) effects. It's just an approximation. This approximation trades away flexibility for simplicity: instead of having to specify the probability of all possible outcomes between 0 and 140, we only have to specify two numbers: $N = 140$ and $p = 0.09$, the parameters of the binomial distribution. These parameters then determine the probabilities for all events in the sample space.

In light of this trade-off, any attempt to draw conclusions from a parametric probability model should also involve the answer to two important questions. First, what unrealistic simplifications have we made in building the model? Second, have these assumptions made our model *too* simple? This second answer will always be context dependent, and it's hard to provide general guidelines about what "too simple" means. Often this boils down to the question of what might go wrong if we use a simplified model, rather than invest the extra work required to build a more complicated model. This is similar to the trade-off that engineers face when they build simplified physical models of something like a suspension bridge or a new fighter jet. Like many things in statistics and probability modeling, this is a case where there is just no substitute for experience and subject-area knowledge.

The expected value

Recall our definition of expected value. Suppose that the possible outcomes for a random variable X are the numbers $x_1, \ldots, x_N$.

Back in Equation 17.1 on page 184, we learned that the formal definition for the expected value of X is

$$E(X) = \sum_{K=1}^{N} P(X = x_i) \cdot x_i \, .$$

Thus the expected value is the probability-weighted average of possible outcomes.

Now let's imagine that X is a binomial random variable: $X \sim \text{Binomial}(N, P)$. If we apply the formal definition of expected value and churn through the math, we find that

$$\begin{aligned} E(X) &= \sum_{k=0}^{k=N} \binom{N}{k} \, P^k \, (1-P)^{N-k} \cdot k \\ &= NP \, . \end{aligned}$$

We've skipped a lot of algebra steps here, but the punchline is a lot more important than the derivation: a random variable with a binomial distribution has expected value $\text{E}(X) = NP$.

This is sometimes called the "NP rule" for expected value: the expected number of events is the number of times an event has to occur (N), times the probability that the event will occur each time (P). The NP rule is a valid way of calculating an expected value precisely for those random events that can be described by a binomial distribution—that is, those events satisfying criteria (1)-(3) on page 186. For random events that *don't* meet these criteria, you'll need to use the formal definition from Equation 17.1 on page 184.

Note: a similar calculation shows that a random variable with a binomial distribution has standard deviation $\text{sd}(X) = \sqrt{NP(1-P)}$.

Advanced optional topic: a derivation of the binomial distribution

To motivate the idea of the binomial distribution, suppose we flip a fair coin only twice.[3] Let our random variable X be the number of times we see "heads" in two coin flips. Thus our sample space for X has three possible outcomes—zero, one, or two. Since the coin flips are independent, all four possible sequences for the two flips (HH, HT, TH, TT) are equally likely, and the probability distribution for X is given by the table below.

The logic of this simple two-flip case can be extended to the general case of N flips: by accounting for every possible sequence of heads and tails that could arise from N flips of a fair coin. Since

[3] By fair, we mean that coin is equally likely to come up heads or tails when flipped.

x_k	$P(X=k)$	Cases
0	0.25	0 heads (TT)
1	0.50	1 head (HT or TH)
2	0.25	2 heads (HH)

successive flips are independent, every sequence of heads and tails has the same probability: $1/2^N$. Therefore,

$$P(X = k \text{ heads}) = \frac{\text{Number of sequences with } k \text{ heads}}{\text{Total number of possible sequences}} . \qquad (17.3)$$

There are 2^N possible sequences, which gives us the denominator. To compute the numerator, we must count the number of these sequences where we see exactly k heads.

How many such sequences are there? To count them, imagine distributing the k heads among the N flips, like putting k items in N boxes, or handing out k cupcakes among N people who want one. Clearly there are N people to which we can assign the first cupcake. Once we've assigned the first, there are $N-1$ people to which we could assign the second cupcake. Then there are $N-2$ choices for the third, and so forth for each successive cupcake. Finally for the kth and final cupcake, there are $N-k+1$ choices. Hence we count

$$N \times (N-1) \times (N-2) \times \cdots \times (N-k+1) = \frac{N!}{(N-k)!}$$

possible sequences, where $N!$ is the factorial function. For example, if $m = 10$ and $k = 7$, this gives 604,800 sequences.

But this is far too many sequences. We have violated an important principle of counting here: don't count the same sequence more than once. The problem is that have actually counted all the ordered sequences, even though we were trying to count unordered sequences. For example, in the $N = 10$, $k = 7$ case, we have counted "Heads on flips $\{1,2,3,4,5,6,7\}$" and "Heads on flips $\{7,6,5,4,3,2,1\}$" as two different sequences. But they clearly both correspond to the same sequence: HHHHHHHTTT.

So how many times have we overcounted each unordered sequence in our tally of the ordered ones? The way to compute this is to count the number of ways we could order k objects. Given a group of k numbers which will be assigned to the "heads" category, we could have chosen from k of the objects to be first in line, from $k-1$ of them to be second in line, from $k-2$ of them to be third in line, and so forth. This means we have counted each unordered sequence $k!$ times. Thus the correct number of ways we could choose k objects out of N possiblities is

$$\frac{N!}{k!(N-k)!} = \binom{N}{k} .$$

For $N = 10$ and $k = 7$, this is 120 sequences—the right answer, and a far cry from the 604,800 we counted above.

Putting all these pieces together, we find that the probability of getting k heads in N flips of a fair coin is

$$P(k \text{ heads}) = \frac{N!}{k!(N-k)!} \, \frac{1}{2^N} = \binom{N}{k} \frac{1}{2^N} . \qquad (17.4)$$

The general case. The above derivation assumes that "yes" (success) and "no" (failure) events are equally likely. Let's now relax this assumption to see where the general definition of the binomial distribution comes from, when the probability of any individual success is not 0.5, but some rather some generic probability p.

Let's take a sequence of N trials where we observed k successes. Each success happens with probability p, and there are k of them. Each failure happens with probability $1 - p$, and there are $m - k$ of them. Because each trial is independent, we multiply all of these probabilities together to get the probability of the whole sequence: $p^k \, (1-p)^{m-k}$. Moreover, our analysis above shows that there are precisely $\binom{N}{k}$ such sequences (i.e. unique ways of getting exactly k successes and $N - k$ failures).

So if we let X denote the (random) number of successes in N trials, then for any value of k from 0 to N,

$$P(X = k) = \binom{N}{k} \, p^k \, (1-p)^{N-k} ,$$

which is the probability mass function given in Equation 17.2.

The Poisson distribution

Our second example of a parametric probability model is the Poisson distribution, named after the French mathematician Siméon Denis Poisson.[4] The Poisson distribution is used to model the number of times than some event occurs in a pre-specified interval of time. For example:

[4] The French speakers among you, or at least the fans of Disney movies, might recognize the word poisson from a different context. Run, Sebastian!

(1) How many goals will Arsenal score in their game against Man U? (The event is a goal, and the interval is a 90-minute game.)

(2) How many couples will arrive for dinner at a hip new restaurant between 7 and 8 PM on a Friday night? (The event is the arrival of a couple asking to sit at a table for two, and the interval is one hour).

(3) How many irate customers will call the 1-800 number for AT&T customer service in the next minute? (The event is a phone call that must be answered by someone on staff, and the interval is one minute.)

In each case, we identify the random variable X as the total number of events that occur in the given interval. The Poisson distribution will provide an appropriate description for this random variable if the following criteria are met:

(1) The events occur independently; seeing one event neither increases nor decreases the probability that a subsequent event will occur.

(2) Events occur the same average rate throughout the time interval. That is, there is no specific sub-interval where events are more likely to happen than in other sub-intervals. For example, this would mean that if the probability of Arsenal scoring a goal in a given 1-minute stretch of the game is 2%, then the probability of a goal during *any* 1-minute stretch is 2%.

(3) The chance of an event occuring in some sub-interval is proportional to the length of that sub-interval. For example, this would mean that if the probability of Arsenal scoring a goal in any given 1-minute stretch of the game is 2%, then the probability that they score during a 2-minute stretch is 4%.

A random variable X meeting these criteria is said to follow a Poisson distribution. The sample space of a Poisson distribution is the set of non-negative integers $0, 1, 2$, etc. This is one important way in which the Poisson differs from the binomial. A binomial random variable cannot exceed N, the number of trials. But there is no fixed upper bound to a Poisson random variable.

The probability mass function of Poisson distribution takes the following form:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \, ,$$

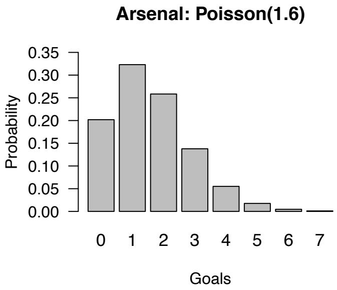
with a single parameter λ (called the rate). This parameter governs the average number of events in the interval: $E(X) = \lambda$. It also governs the standard deviation: $\text{sd}(X) = \sqrt{\lambda}$.

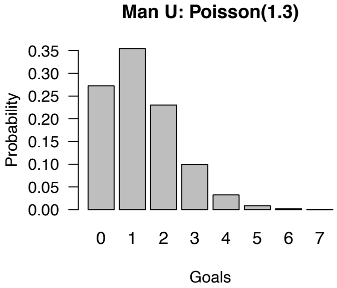
Example: modeling the score in a soccer game. Let's return to our soccer game example. Across all games in the 2015-16 English Premiere League (widely considered to be the best professional soccer league in the world), Arsenal scored 1.6 goals per game, while Manchester United scored 1.3 goals per game. How likely is it that Arsenal beats Man U? How likely is a scoreless draw at 0-0? To answer these questions, let's make some simplifying assumptions.

(1) Let X_A be the number of goals scored in a game by Arsenal. We will assume that X_A can be a described by a Poisson distribution with rate parameter 1.6: that is, $X_A \sim \text{Poisson}(\lambda = 1.6)$.

(2) Let X_M be the number of goals scored in a game by Manchester United. We will assume that $X_M \sim \text{Poisson}(\lambda = 1.3)$.

(3) Finally, we will assume that X_A and X_M are independent of one another.





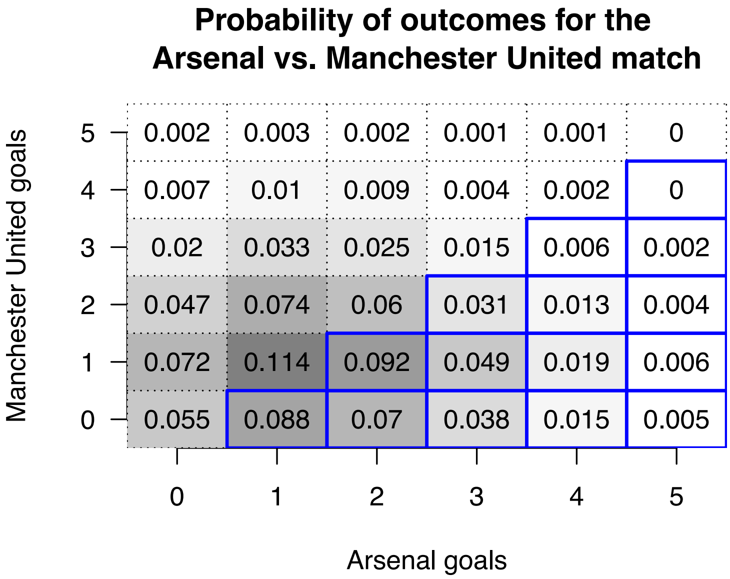
Our model sets the rate parameters for each team's Poisson distribution to match their average scoring rates across the season. The corresponding PMFs are shown at right.

Under these simplifying assumptions, we can calculate the probability of any possible score—for example, Arsenal 2–0 Manchester United. Because we have assumed that X_A and X_M are independent, we can multiply together the two probabilities we get from each random variable's Poisson distribution:

$$P(X_A = 2, X_M = 0) = \left(\frac{1.6^2}{2!} e^{-1.6} \right) \cdot \left(\frac{1.3^0}{0!} e^{-1.3} \right) \approx 0.07\,.$$

Figure 17.3 shows a similar calculation for all scores ranging from 0–0 to 5–5 (according to the model, the chance of a score larger than this is only 0.6%). By summing up the probabilities for the various score combinations, we find that:

- Arsenal wins with probability 44%.
- Man U wins with probability 31%.
- The game ends in a draw with probability 25%. In particular, a scintillating 0–0 draw happens with probability 5.5%.



Manchester United goals \ Arsenal goals	0	1	2	3	4	5
5	0.002	0.003	0.002	0.001	0.001	0
4	0.007	0.01	0.009	0.004	0.002	0
3	0.02	0.033	0.025	0.015	0.006	0.002
2	0.047	0.074	0.06	0.031	0.013	0.004
1	0.072	0.114	0.092	0.049	0.019	0.006
0	0.055	0.088	0.07	0.038	0.015	0.005

Figure 17.3: A matrix of probabilities associated with various match scores under the independent Poisson model of an Arsenal vs. Man U match, based on scoring statistics from 2015-16 Premiere League season. Each entry in the matrix is the probability with the corresponding score (darker grey = higher probability). The cells outlined in blue correspond to an Arsenal win, which happens with probability 44% (versus 25% for a draw and 31% for a Manchester United win.

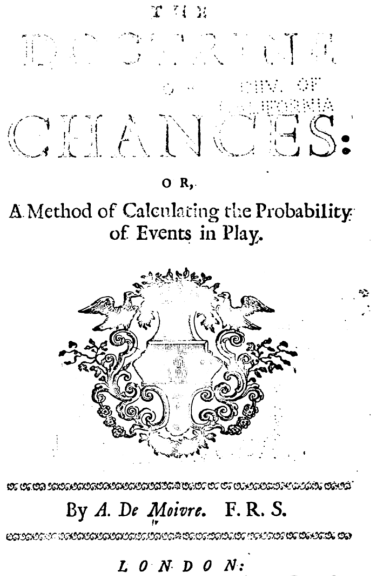
The normal distribution

This chapter's third and final example of a parametric probability model is the normal distribution—the most famous and widely used distribution in the world.

Some history

Historically, the normal distribution arose an an approximation to the binomial distribution. In 1711, a Frenchman named Abraham de Moivre published a book called *The Doctrine of Chances*. The book was reportedly was prized by gamblers of the day for its many useful calculations that arose in dice and card games. In the course of writing about these games, de Moivre found it necessary to perform computations using the binomial distribution for very large values of N, the number of independent trial in a binomial distribution. (Imagine flipping a large number of coins and making bets on the outcomes, and you too will see the necessity of this seemingly esoteric piece of mathematics.)

As you recall the previous section, these calculations require computing binomial coefficients $\binom{N}{k}$ for very large values of N. But because these computations involve the factorial function,


THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability of Events in Play.

By A. De Moivre. F. R. S.

LONDON:

Figure 17.4: The title page of de Moivre's "The Doctrine of Chances" (1711), from an early edition owned by the University of California, Berkeley. One interesting thing about the history of statistics is the extent to which beautiful mathematical results came out of the study of seemingly trivial gambling

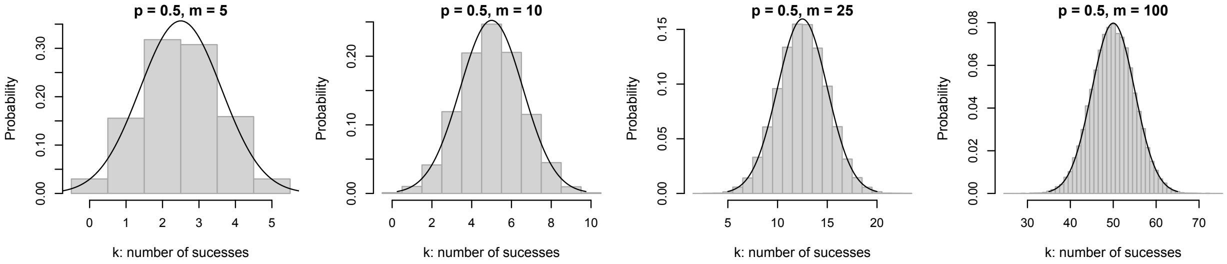


Figure 17.5: The binomial distribution for $p = 0.5$ and an increasingly large number of trials, together with de Moivre's normal approximation.

they were far too time-consuming without modern computers, which de Moivre didn't have. So he derived an approximation based on the number $e \approx 2.7183$, the base of the natural logarithm. He discovered that, if a random variable X has a binomial distribution with parameters N and p, which we recall is written $X \sim \text{Binomial}(N, p)$, then the approximate probability that $X = k$ is

$$P(X = k) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{(k-\mu)^2}{2\sigma^2}} \,, \tag{17.5}$$

where $\mu = mp$ and $\sigma^2 = Np(1 - p)$ are the expected value and variance, respectively, of the binomial distribution. When considered as a function k, this results in the familiar bell-shaped curve plotted in Figure 17.5—the famous *normal distribution*.

We can usually (though not always) avoid working with this expression directly, since every piece of statistical software out there can compute probabilities under the normal distribution. The important thing to notice is how the binomial samples in Figure 17.5 start to look more normal as the number of trials N gets progressively larger: first 5, then 10, 25, and finally 100. The histograms show the binomial distribution itself, while the black curves show de Moivre's approximation. Clearly he was on to something. This famous result of de Moivre's is usually thought of as the first *central limit theorem* in the history of statistics, where the word "central" should be understood to mean "fundamental."

The normal distribution: a modern understanding

The other term for the normal distribution is the Gaussian distribution, named after the German mathematician Carl Gauss. This raises a puzzling question. If de Moivre invented the normal approximation to the binomial distribution in 1711, and Gauss (1777–1855) did his work on statistics almost a century after de Moivre,

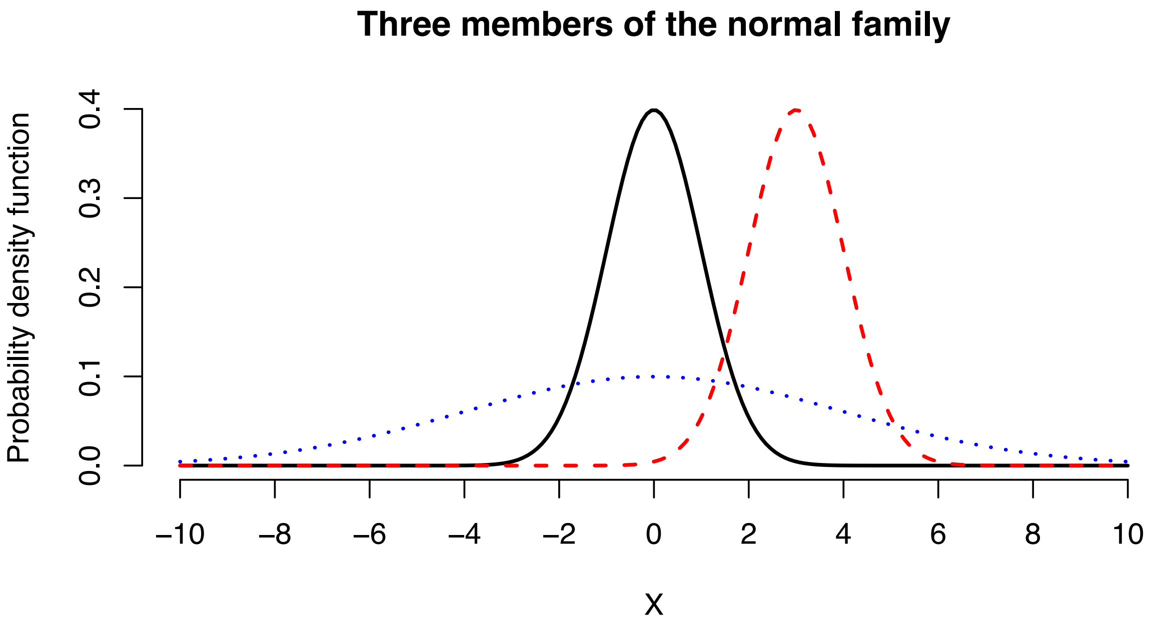


Figure 17.6: Three members of the normal family: $\mathrm{N}(0,1^2)$, $\mathrm{N}(0,4^2)$, and $\mathrm{N}(3,1^2)$. See if you can identify which is which using the guideline that 95% of the probability will be within two standard deviations σ of the mean. Remember, the second parameter is the variance σ^2, not the standard deviation. So $\sigma^2 = 4^2$ means a variance of 16 and a standard deviation of 4.

why then is the normal distribution also named after Gauss and not de Moivre? This quirk of eponymy arises because de Moivre only viewed his approximation as a narrow mathematical tool for performing calculations using the binomial distribution. He gave no indication that he saw it as a more widely applicable probability distribution for describing random phenomena. But Gauss—together with another mathematician around the same time, named Laplace—did see this, and much more.

If we want to use the normal distribution to describe our uncertainty about some random variable X, we write $X \sim \mathrm{N}(\mu, \sigma^2)$. The numbers μ and σ^2 are parameters of the distribution. The first parameter, μ, describes where X tends to be centered; it also happens to be the expected value of the random variable. The second parameter, σ^2, describes how spread out X tends to be around its expected value; it also happens to be the variance of the random variable. Together, μ and σ^2 completely describe the distribution, and therefore completely characterize our uncertainty about X.

The normal distribution is described mathematically by its probability density function, or PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) . \qquad (17.6)$$

If you plot this as a function of x, you get the famous bell curve (Figure 17.6). How can you interpret a "density function" like this one? If you the take the area under this curve between two values z_1 and z_2, you will get the probability that the random variable X will end up falling between z_1 and z_2 (see Figure 17.7). The height of the curve itself is a little more difficult to interpret, and we won't worry about doing so—just focus on the "area under the curve" interpretation.

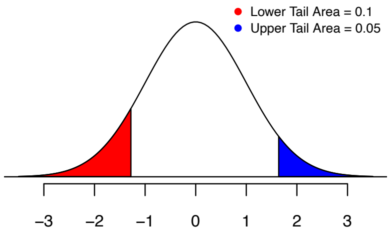


Figure 17.7: Examples of upper and lower tail areas. The lower tail area of 0.1 is at $z = -1.28$. The upper tail area of 0.05 is at $z = 1.64$

Here are two useful facts about normal random variables areas—or more specifically, about the central areas under the curve, between the tails. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then the chance that X will be within 1σ of its mean is about 68%, and the chance that it will be within 2σ of its mean is about 95%. Said in equations:

$$\begin{aligned} P(\mu - 1\sigma < X < \mu + 1\sigma) &\approx 0.68 \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &\approx 0.95 . \end{aligned}$$

Actually, it's more like 1.96σ rather than 2σ for the second part. So if your problem requires a level of precision to an order of 0.04σ or less, then don't use this rule of thumb, and instead go with the true multiple of 1.96.

When is the normal distribution an appropriate model?

The normal distribution is now used as a probability model in situations far more diverse than de Moivre, Gauss, or Laplace ever would have envisioned. But it still bears the unmistakeable traces of its genesis as a large-sample approximation to the binomial distribution. That is, it tends to work best for describing situations where each normally distributed random variable can be thought of as the sum of many tiny, independent effects of about the same size, some positive and some negative. In cases where this description doesn't apply, the normal distribution may be a poor model of reality. Said another way: the normal distribution describes an aggregation of nudges: some up, some down, but all pretty small.

As a result, the normal distribution shares the property of the binomial distribution that huge deviations from the mean are unlikely. It has, in statistical parlance, "thin tails." Using our rule

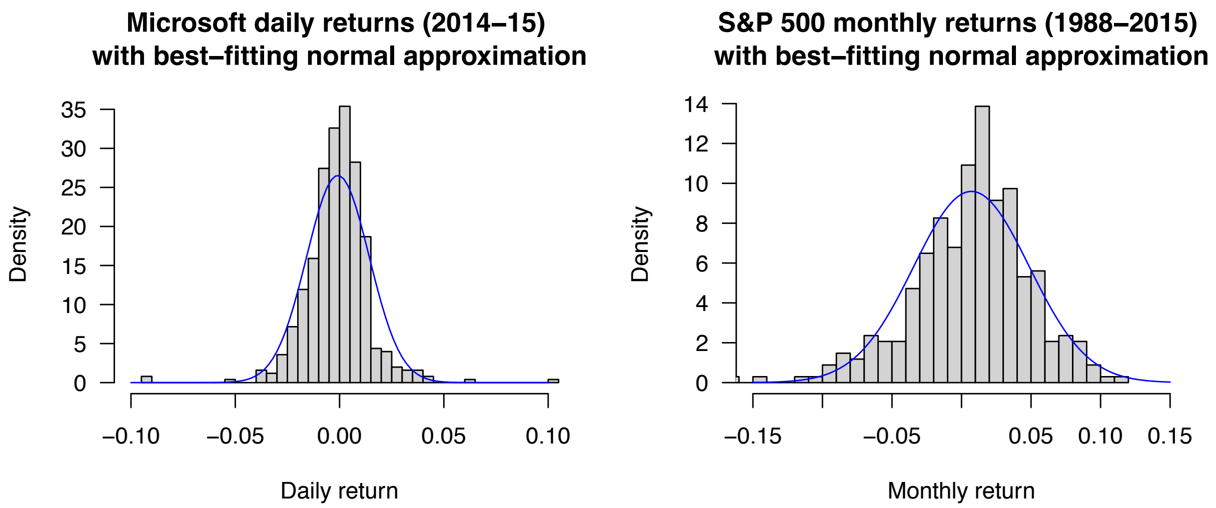


Figure 17.8: Daily stock returns for Microsoft (left) and the S&P 500 (right), together with the best-fitting normal approximations. The approximation on the right is not bad, while the approximation on the left drastically underestimates the probability of extreme results.

of thumb above, a normally distributed random variable has only a 5% chance of being more than two standard deviations away from the mean. It also has less than a 0.3% chance of being more than three standard deviations away from the mean. Large outliers are vanishingly rare.

For example, in the histogram of daily returns for Microsoft stock in the left panel Figure 17.8, notice the huge outliers in the lower tail. These returns would be wildly implausible if the returns really followed a normal distribution. A daily return tends to be dominated by one or two major pieces of information. It does not resemble an aggregation of many independent up-or-down nudges, and so from first principles alone, we should probably expect the normal distribution to provide a poor fit. As we would expect, the best-fitting normal approximation (i.e. the one that matches the sample mean and standard deviation of the data) does not fit especially well.

The example of Microsoft stock recalls the earlier discussion on the trustworthiness of the simplifying assumptions that must go into building a probability model. To recap:

> Have these assumptions made our model too simple? This . . . answer will always be context dependent, and it's hard to provide general guidelines about what "too simple" means. Often this boils down to the questin of what might go wrong if we use a simplified model, rather than invest the extra work

required to build a more complicated model.

What might go wrong if we use a normal probability model for Microsoft returns? In light of what we've seen here, the answer is: we might be very unpleasantly surprised by monetary losses that are far more extreme than envisioned under our model. This sounds very bad, and is probably a sufficient reason not to use the normal model in the first place. To make this precise, observe that the 2 most extreme daily returns for Microsoft stock were both 6 standard deviations below the mean. According to the normal model, we should only expect to see such an extreme result once every billion trading days, since

$$P(X < \mu - 6\sigma) \approx 10^{-9}\,.$$

This is a wildly overtoptimistic assessment, given that we actually saw two such results in the 503 trading days from 2014-15.

On the other hand, the normal distribution works a lot better for stock indices than it does for individual stocks, especially if we aggregate those returns over a month rather than only a day, so that the daily swings tend to average out a bit more. Take, for example, the best-fitting normal approximation for the monthly returns of the S&P 500 stock index from 1988 to 2015, in the right panel of Figure 17.8. Here the best-fitting normal distribution, though imperfect, looks a lot better than the corresponding fit for an individual stock on the left. Here, the most extreme monthly return was 4 standard deviations below the mean (which happened in October 2008, during the financial crisis of that year that augured the Great Recession). According to the normal model, we would expect such an extreme event to happen with about 2% probability in any given 27-year stretch. Thus our model looks a tad optimistic, but not wildly so.

Example: modeling a retirement portfolio

From 1900–2015, the average annual return[5] of the S&P 500 stock index is 6.5%, with a standard deviation of 19.6%. Let's use these facts to build a probability model for the future 40-year performance of a $10,000 investment in a diversified portfolio of U.S. stocks (i.e. an index fund). While there's no guarantee that past returns are a reliable guide to future returns, they're the only data we have. After all, as Mark Twain is reputed to have said, "History doesn't repeat itself, but it does rhyme."

[5] Real returns net of infation and dividends. Remember that a return is simply the implied interest rate from holding an asset for a specified period. If you buy a stock at $100 and sell a year later at $110, then your return is $(110 - 100)/100 = 0.1$, or 10%. If inflation over that year was 3%, then your real return was 7%.

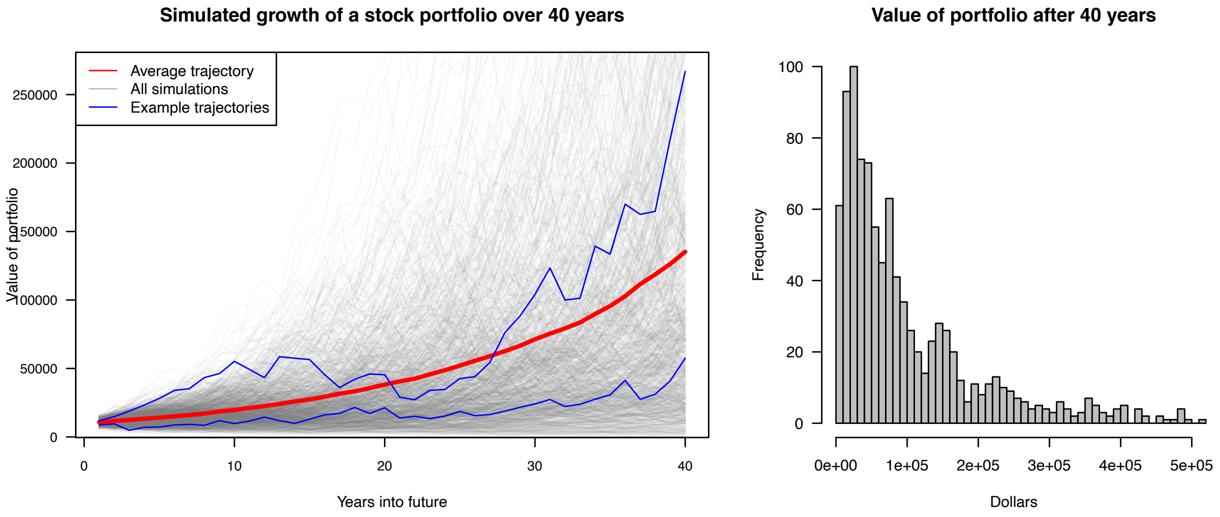


Figure 17.9: Left panel: 1000 simulated trajectories for the growth of a $10,000 stock investment over 40 years, assuming that year stock returns are normally distributed with a mean of 6.5% and a standard deviation of 19.6%. Two individual trajectories (leading to very different outcomes) are highlighted in blue; the average trajectory is shown in red. The right panel shows the simulated probability distribution for W_{40}, the final value of the portfolio after 40 years of random returns.

Let's say that your initial investment is $W_0 = \$10,000$, and that X_t is the return of your portfolio in year t expressed as a decimal fraction (e.g. a 10% return in year 1 would mean that $X_t = 0.1$). Here t will run from 1 to 40, since we want to track your portfolio over 40 years. If we knew the returns $X_1, X_2, \ldots, X_{40}$ all the way into the future, we could calculate your terminal wealth as

$$W_{40} = W_0 \cdot \prod_{t=1}^{40} (1 + X_t) \, ,$$

by simply compounding the interest year after year.[6] This formula follows from the fact that W_{t+1}, your wealth in year t, is given by the simple interest formula: $W_{t+1} = W_t \cdot (1 + X_t)$. Accumulating returns year after year then gives us the above formula.

Of course, we don't know these interest rates. But we do have a probability model for them, whose parameters have been chosen to match the historical record: $X_t \sim N(\mu = 0.065, \sigma^2 = 0.196^2)$. Thus to estimate the probability distribution of the random variable W_{40}, your terminal wealth after 40 years, we will use a Monte Carlo simulation, in which we repeat the following steps many thousands of times:

(1) Simulate random returns from the normal probability model: $X_t \sim N(0.065, 0.196^2)$ for $t = 1, \ldots, 40$.

(2) Starting with year $t = 1$ and ending with year $t = 40$,

[6] Here the symbol $\prod$ means we take the running product of all the terms, from $t = 1$ to $t = 40$, just like Σ means we take a running sum.

chain these simulated interest rates together using the simple-interest formula

$$W_{t+1} = W_t \cdot (1 + X_t)$$

to form a single simulated trajectory $W_1, W_2, \ldots, W_{40}$ of wealth.

As a byproduct of this, we get a simulated probability distribution of W_t for all values of t from 1 up to 40.

Figure 17.9 shows 1000 trajectories simulated according to this algorithm, along with the histogram of the 1000 different values of W_{40}, your wealth in 40 years. There are several interesting things to point out about the result:

(1) The *average* trajectory in Figure 17.9 results in a final value of $W_{40} \approx \$135{,}000$ from your initial \$10,000 investment.[7]

(2) But there is tremendous variability about this average trajectory, both over time for a single trajectory, and across all trajectories. To illustrate this point, two simulated trajectories are shown in blue in Figure 17.9: one resulting in a final portfolio of about \$250,000, and another resulting in less than \$50,000.

(3) The simulated probability distribution of final wealth (right panel of Figure 17.9) was constructed using nothing but normally distributed random variables as inputs. But this distribution is itself highly non-normal.[8] This provides a good example of using Monte Carlo simulation to simulate a complex probability distribution by breaking down into a function of many smaller, simpler parts (in this case, the yearly returns).

(4) The estimated probability that your \$10,000 investment will have lost money (net of inflation) after 10 years is about 19%; after 20 years, about 13%; after 40 years, about 6%.

(5) The estimated probability that your investment will grow to \$1 million or more after 40 years is about 1%.

The moral of the story is that the stock market is probably a good way to get rich over time. But there's a nonzero chance of losing money—and the riches come only in the long run, and with a lot of uncertainty about how things will unfold along the way.

Postscript

We've now seen three examples of parametric probability models: a binomial model for airline no-shows, a Poisson model for

[7] Remember that our assumed rates of return are adjusted for inflation, so this corresponds to the purchasing power of \$135,000 in today's money. The actual dollar value of this portfolio, as measured in the currency of the future, would be a good deal higher.

[8] In particular it has a long right tail, reflecting the small probability of explosive growth in your investment.

scoring in a soccer game, and a normal model for annual returns of the stock market. In each case, we chose the parameters of the probability model from real-world data, using simple and obvious criteria (e.g. the overall no-show rate for commercial flights, or the mean return of stocks over the last century).[9] In essence, we performed a naïve form of statistical inference for the parameters of our probability models. This intersection where probability modeling meets data is an exciting place where the big themes of the book all come together.

[9] Technically what we did here was called *moment matching,* wherein we match sample moments (e.g. mean, variance) of the data to the corresponding moments of the probability distribution.