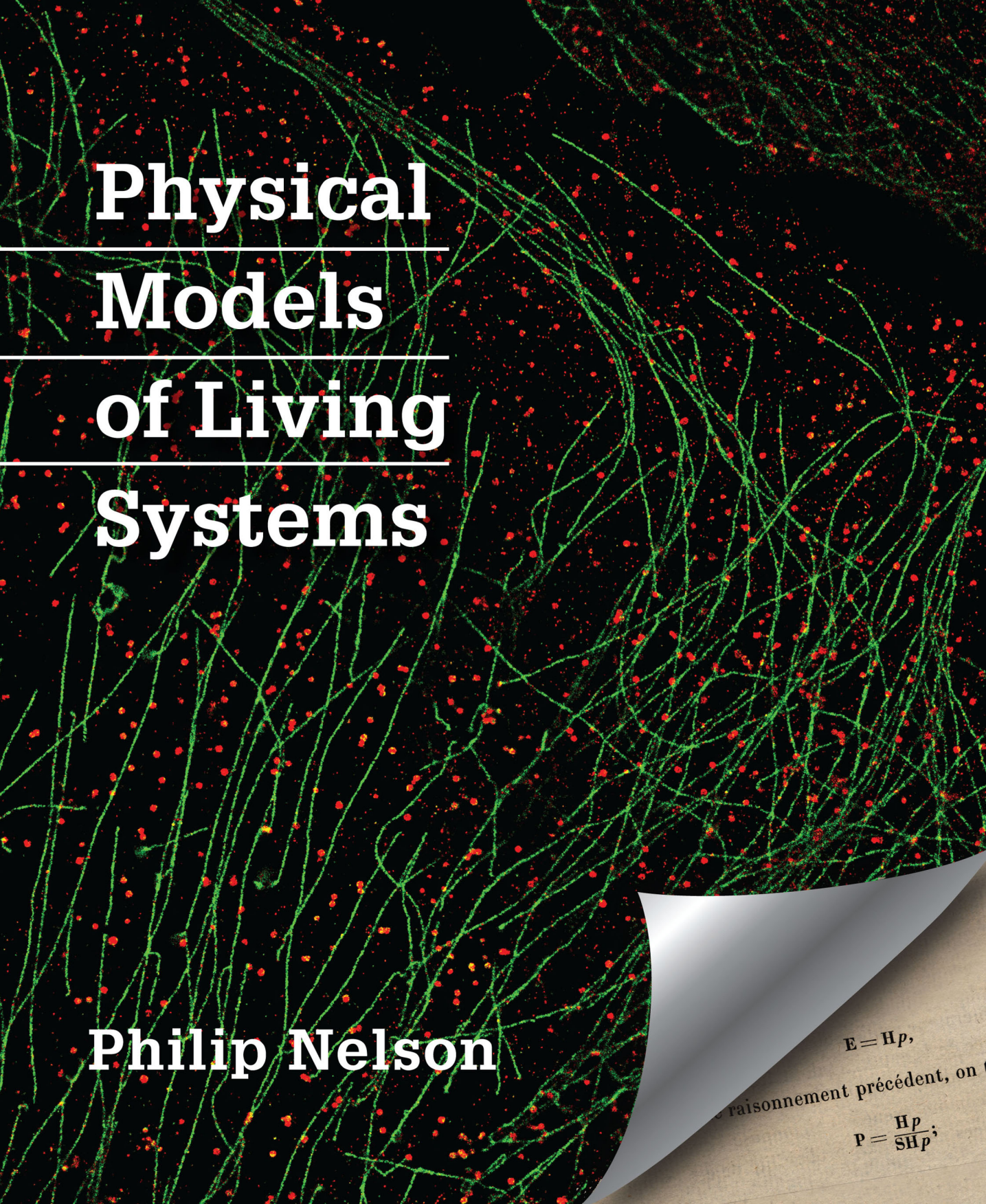
Physical Models of Living Systems

Philip Nelson

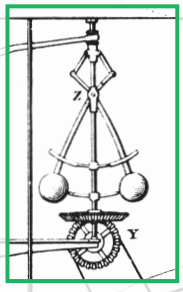


Physical Models of Living Systems

Philip Nelson

University of Pennsylvania

with the assistance of Sarina Bromberg, Ann Hermundstad, and Jason Prentice



W. H. Freeman & Company

A Macmillan Education Imprint

Publisher: Kate Parker
Acquisitions Editor: Alicia Brady
Senior Development Editor: Blythe Robbins
Assistant Editor: Courtney Lyons
Editorial Assistant: Nandini Ahuja
Marketing Manager: Taryn Burns
Senior Media and Supplements Editor: Amy Thorne
Director of Editing, Design, and Media Production: Tracey Kuehn
Managing Editor: Lisa Kinne
Project Editor: Kerry O'Shaughnessy
Production Manager: Susan Wein
Design Manager and Cover Designer: Vicki Tomaselli
Illustration Coordinator: Matt McAdams
Photo Editors: Christine Buese, Richard Fox
Composition: codeMantra
Printing and Binding: RR Donnelley

Cover: [Two-color, superresolution optical micrograph.] Two specific structures in a mammalian cell have been tagged with fluorescent molecules via immunostaining: microtubules (false-colored *green*) and clathrin-coated pits, cellular structures used for receptor-mediated endocytosis (false-colored *red*). See also Figure 6.5 (page 138). The magnification is such that the height of the letter "o" in the title corresponds to about 1.4 μm. [Image courtesy Mark Bates, Dept. of NanoBiophotonics, Max Planck Institute for Biophysical Chemistry, published in Bates et al., 2007. Reprinted with permission from AAAS.] *Inset:* The equation known today as the "Bayes formula" first appeared in recognizable form around 1812, in the work of Pierre Simon de Laplace. In our notation, the formula appears as Equation 3.17 (page 52) with Equation 3.18. (The letter "S" in Laplace's original formulation is an obsolete notation for sum, now written as $\sum$.) This formula forms the basis of statistical inference, including that used in superresolution microscopy.
Title page: Illustration from James Watt's patent application. The green box encloses a centrifugal governor. [From *A treatise on the steam engine: Historical, practical, and descriptive* (1827) by John Farey.]

Library of Congress Preassigned Control Number: 2014949574
ISBN-13: 978-1-4641-4029-7
ISBN-10: 1-4641-4029-4

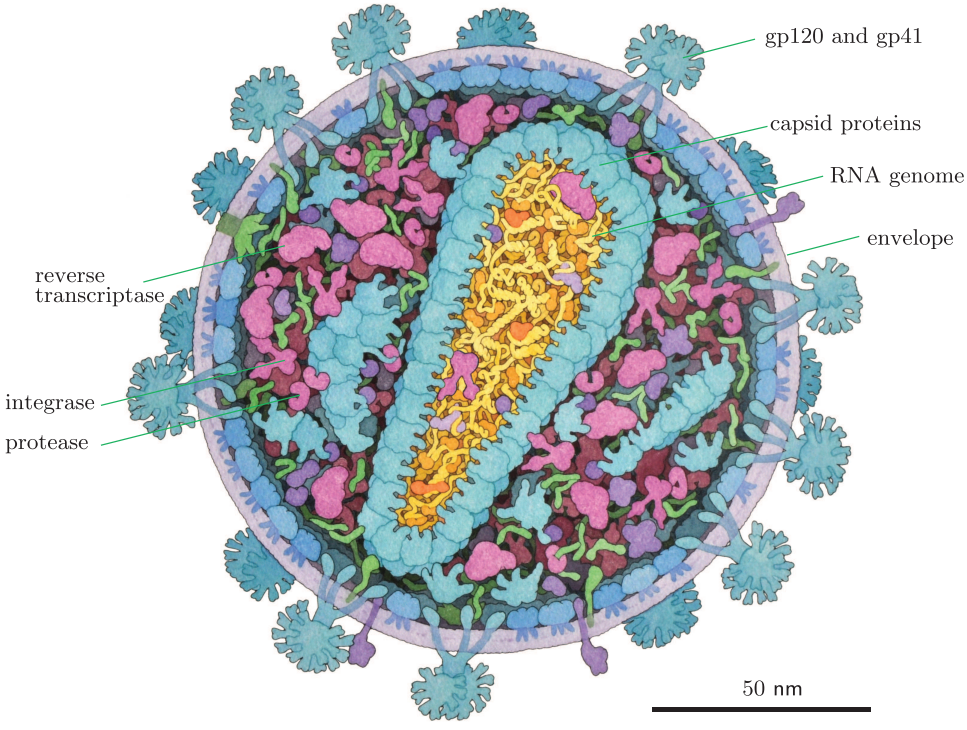
Printed in the United States of America

First printing

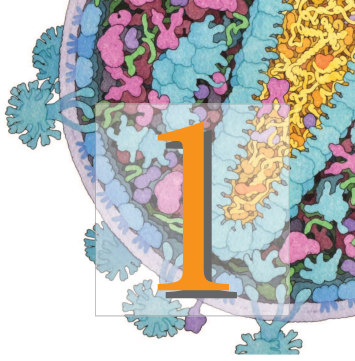
W. H. Freeman and Company, 41 Madison Avenue, New York, NY 10010
Houndmills, Basingstoke RG21 6XS, England
www.whfreeman.com

PART I

First Steps



[Artist's reconstructions based on structural data.] **A human immunodeficiency virus particle** (virion), surrounded by its lipid membrane envelope. The envelope is studded with gp120, the protein that recognizes human T cells. The envelope encloses several enzymes (proteins that act as molecular machines), including HIV protease, reverse transcriptase, and integrase. Two RNA strands carrying the genome of HIV are packaged in a cone-shaped protein shell called the capsid. See also **Media 1.** [Courtesy David S Goodsell.]



1 Virus Dynamics

We all know that Art is not truth. Art is a lie that makes us realize the truth.
—Pablo Picasso

1.1 First Signpost

The Prolog suggested a three-step procedure to make headway on a scientific problem (see page 4). Unfortunately, the experiment that can be performed usually does not directly yield the information we desire, and hence does not directly confirm or disprove our original hypothesis. For example, this chapter will argue that testing the viral mutation hypothesis in the Prolog actually requires information not directly visible in the data that were available in 1995.

Thus, a fourth step is almost always needed:

4. Embody the physical metaphor (or **physical model**) in mathematical form, and attempt to fit it to the experimental data.

In this statement, **fit** means "adjust one or more numbers appearing in the model." For each set of these **fit parameter** values that we choose, the model makes a prediction for some experimentally measurable quantity, which we compare with actual observations. If a successful fit can be found, then we may call the model "promising" and begin to draw tentative conclusions from the parameter values that yield the best fit. This chapter will take a closer look at the system discussed in the Prolog, illustrating how to construct a physical model, express it in mathematical form, fit it to data, evaluate the adequacy of the fit, and draw conclusions. The chapter will also get you started with some of the basic computer skills needed to carry out these steps.

Each chapter of this book begins with a biological question to keep in mind as you read, and an idea that will prove relevant to that question.
This chapter's Focus Question is
Biological question: Why did the first antiviral drugs succeed briefly against HIV, but then fail?
Physical idea: A physical model, combined with a clinical trial designed to test it, established a surprising feature of HIV infection.

1.2 Modeling the Course of HIV Infection

We begin with just a few relevant facts about HIV, many of which were known in 1995. It will not be necessary to understand these in detail, but it is important to appreciate just how much was already known at that time.

1.2.1 Biological background

In 1981, the US Centers for Disease Control noticed a rising incidence of rare diseases characterized by suppression of the body's immune system. As the number of patients dying of normally nonlethal infections rose, it became clear that some new disease, with an unknown mechanism, had appeared. Eventually, it was given the descriptive name acquired immune deficiency syndrome (AIDS).

Two years later, research teams in France and the United States showed that a virus was present in lymph fluid taken from AIDS patients. The virus was named human immunodeficiency virus (HIV). To understand why HIV is so difficult to eradicate, we must very briefly outline its mechanism as it was later understood.

HIV consists of a small package (the virus particle, or **virion**) containing some nucleic acid (two copies of the genome), a protective protein shell (or **capsid**), a few other protein molecules needed for the initial steps of infection, and a surrounding **envelope** (see the figure on page 7). In a **retrovirus** like HIV, the genome takes the form of RNA molecules, which must be converted to DNA during the infection.[1]

The genome of HIV is extremely short, roughly 10 000 bases. It contains nine genes, which direct the synthesis of just 19 different proteins. Three of the genes code for proteins that perform the following functions:[2]

- *gag* generates the four proteins that make the virion's capsid.
- *pol* generates three protein machines (enzymes): A **reverse transcriptase** converts the genome to DNA, an **integrase** helps to insert the viral DNA copy into the infected cell's own genome, and a **protease** cleaves (cuts) the product of *gag* (and of *pol* itself) into separate proteins (Figure 1.1).
- *env* generates a protein that embeds itself in the envelope (called gp41), and another (called gp120; see page 7) that attaches to gp41, protrudes from the envelope, and helps the virus to target and enter its host.

HIV targets some of the very immune cells that normally protect us from disease. Its gp120 protein binds to a receptor found on a human immune cell (the **CD4+ helper T cell**,

[1] The normal direction of information transmission is from DNA to RNA; a *retro*virus is so named because it reverses this flow.
[2] That is, they are "structural" genes. The other six genes code for transcription factors; see Chapter 9.

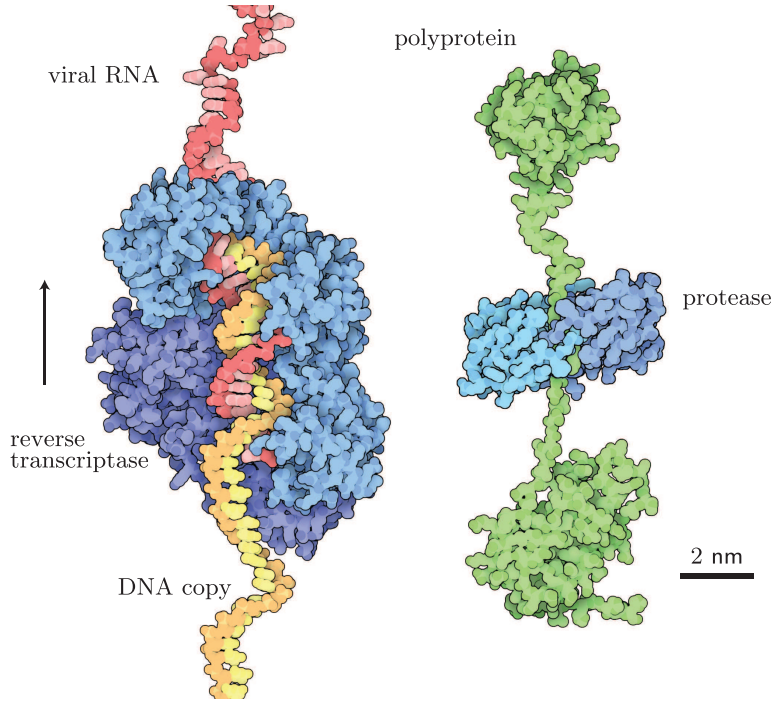


Figure 1.1 [Artist's reconstructions based on structural data.] **Two protein machines needed for HIV replication.** *Left:* Reverse transcriptase is shown transcribing the viral RNA into a DNA copy. This molecular machine moves along the RNA (*arrow*), destroying it as it goes and synthesizing the corresponding DNA sequence. *Right:* HIV protease cleaves the polyprotein generated by a viral gene (in this case, *gag*) into individual proteins. Many antiviral drugs block the action of one of these two enzymes. [Courtesy David S Goodsell.]

or simply "T cell"). Binding triggers fusion of the virion's envelope with the T cell's outer membrane, and hence allows entry of the viral contents into the cell. The virion includes some ready-to-go copies of reverse transcriptase, which converts the viral genome to DNA, and integrase, which incorporates the DNA transcript into the host cell's own genome. Later, this rogue DNA directs the production of more viral RNA. Some of the new RNA is translated into new viral proteins. The rest gets packaged with those proteins, to form several thousand new virions from each infected cell. The new virions escape from the host cell, killing it and spreading the infection. The immune system can keep the infection in check for many years, but eventually the population of T cells falls. When it drops to about 20% of its normal value, then immune response is seriously compromised and the symptoms of AIDS appear.

The preceding summary is simplified, but it lets us describe some of the drugs that were available by the late 1980s. The first useful anti-HIV drug, zidovudine (or AZT), blocks the action of the reverse transcriptase molecule. Other **reverse transcriptase inhibitors** have since been found. As mentioned in the Prolog, however, their effects are generally short lived. A second approach targets the protease molecule; **protease inhibitors** such as ritonavir result in defective (not properly cleaved) proteins within the virion. The effects of these drugs also proved to be temporary.

1.2.2 An appropriate graphical representation can bring out key features of data

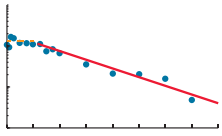


Figure 0.3 (page 4)

Clearly, HIV infection is a complex process. It may seem that the only way to checkmate such a sophisticated adversary is to keep doggedly looking for, say, a new protease inhibitor that somehow works longer than the existing ones.

But sometimes in science the intricate details can *get in the way* of the viewpoint shift that we need for a fresh approach. For example, the Prolog described a breakthrough that came only after appreciating, and documenting, a very basic aspect of HIV infection: its ability to evolve rapidly in a single patient's body.

The Prolog suggested that fast mutation is possible if the viral replication rate is high, and that this rate could be determined by examining the falloff of viral load when production was halted by a drug. The graph in Figure 0.3 is drawn in a way that makes a particular behavior manifest: Instead of equally spaced tick marks representing uniform intervals on the vertical axis, Figure 0.3 uses a logarithmic scale. That is, each point is drawn at a height above the horizontal axis that is proportional to the logarithm of virus population. (The horizontal axis is drawn with an ordinary linear scale.) The resulting **semilog plot** makes it easy to see when a data series is an exponential function of time:[3] The graph of the function $f(t) = C\exp(kt)$ will appear as a straight line, because $\ln f(t) = (\ln C) + kt$ is a linear function of t.

Your Turn 1A

Does this statement depend on whether we use natural or common logarithms?

Log axes are usually labeled at each power of 10, as shown in Figure 0.3. The unequally spaced "minor" tick marks between the "major" ones in the figure are a visual cue, alerting the reader to the log feature. Most computer math packages can create such axes for you. Note that the tick marks on the vertical axis in Figure 0.3 represent 1000, 2000, 3000, . . . , 9000, 10 000, 20 000, 30 000, . . . , 900 000, 1 000 000. In particular, *the next tick after* 10^4 *represents* $2 \cdot 10^4$, *not* 11 000, and so on.

1.2.3 Physical modeling begins by identifying the key actors and their main interactions

Although the data shown in Figure 0.3 are suggestive, they are not quite what we need to establish the hypothesis of rapid virus mutation. The main source of mutations is reverse transcription.[4] Reverse transcription usually occurs only once per T cell infection. So if we wish to establish that there are many opportunities for mutation, then we need to show that many *new T cell infections* are occurring per unit time. This is not the same thing as showing that new virions are being created rapidly, so we must think more carefully about what we can learn from the data. Simply making a graph is not enough.

Moreover, the agreement between the data (dots in Figure 0.3) and the simple expectation of exponential falloff (line) is actually not very good. Close inspection shows that the rapid fall of virus population does not begin at the moment the drug is administered

[3] The prefix "semi-" reminds us that only one axis is logarithmic. A "log-log plot" uses logarithmic scales on *both* axes; we'll use this device later to bring out a different feature in other datasets.

[4] The later step of making new viral genomes from the copy integrated into the T cell's DNA is much more accurate; we can neglect mutations arising at that step.

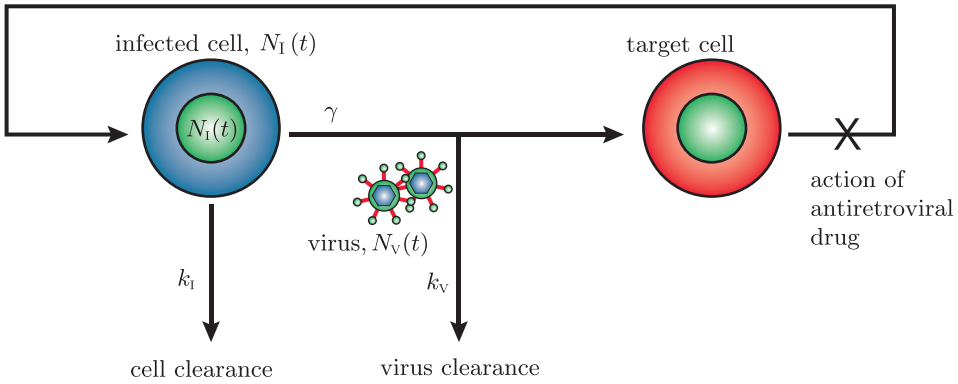


Figure 1.2 [Schematic.] **Simplified virus life cycle.** In this model, the effect of antiviral drug therapy is to halt new infections of T cells (*cross*). The constants k_I, k_V, γ introduced in the text are shown at the points of action of the associated processes.

("time 0"). Instead, the data shown in the figure (and similar graphs from other patients) show an initial pause in virus population at early times, prior to the exponential drop. It's not surprising, really—so far we haven't even attempted to write any quantitative version of our initial intuition.

To do better than this, we begin by identifying the relevant processes and quantities affecting virus population. Infected T cells produce free virions, which in turn infect new cells. Meanwhile, infected T cells eventually die, and the body's defenses also kill them; we will call the combined effect **clearance** of infected cells. The immune system also destroys free virions, another kind of clearance. To include these processes in our model, we first assign names to all the associated quantities. Thus, let t be time after administering the antiviral drug. Let $N_I(t)$ be the number of infected T cells at time t, and $N_V(t)$ the number of free virions in the blood (the **viral load**).[5]

Figure 0.1 (page 1)

Before drug treatment (that is, at time $t < 0$), production and removal processes roughly balance, leading to a nearly steady (**quasi-steady**) state, the long period of low virus population seen in Figure 0.1. In this state, the rate of new infections must balance T cell clearance—so finding their rate of clearance will tell us the rate of new infections in the quasi-steady state, which is what we are seeking.

Let's simplify by assuming that the antiviral drug completely stops new infections of T cells. From that moment, uninfected T cells become irrelevant—they "decouple" from infected T cells and virions. We also simplify by assuming that each infected T cell has some fixed chance of being cleared in any short time interval. That chance depends on the duration Δt of the interval, and it becomes zero if $\Delta t = 0$, so it's reasonable to suppose that it's the product of Δt times a constant, which we'll call k_I.[6] These assumptions imply that, after administering the drug, N_I changes with time according to a simple equation:

[5] The concentration of virions is therefore $N_V(t)$ divided by the blood volume. Because the blood volume is constant in any particular patient, we can work with either concentration or total number.

[6] T2 We are ignoring the possibility of saturation, that is, decrease in clearance probability per infected cell per time when the concentration of infected cells is so high that the immune system is overloaded. This assumption will not be valid in the late stages of infection. We also assume that an infected T cell is unlikely to divide before being cleared.

$$\frac{dN_I}{dt} = -k_I N_I \qquad \text{for } t \geq 0. \tag{1.1}$$

In this formula, the **clearance rate constant** k_I is a **free parameter** of the model—we don't know its value in advance. Notice that it appears multiplied by N_I: The number lost between t and $t + \Delta t$ depends on the total number present at time t, as well as on the rate constant and Δt.

Simultaneously with the deaths of infected T cells, virions are also being produced and cleared. Similarly to what we assumed for N_I, suppose that each virion has a fixed probability per time to be cleared, called k_V, and also that the number *produced* in a short interval Δt is proportional to the population of infected T cells. Writing the number produced as γN_I, where the constant of proportionality γ is another free parameter, we can summarize our physical model by supplementing Equation 1.1 with a second equation:

$$\frac{dN_V}{dt} = -k_V N_V + \gamma N_I. \tag{1.2}$$

Figure 1.2 summarizes the foregoing discussion in a cartoon. For reference, the following quantities will appear in our analysis:

t	time since administering drug
$N_I(t)$	population of infected T cells; its initial value is N_{I0}
$N_V(t)$	population of virions; its initial value is N_{V0}
k_I	clearance rate constant for infected T cells
k_V	clearance rate constant for virions
γ	rate constant for virion production per infected T cell
β	an abbreviation for γN_{I0}

1.2.4 Mathematical analysis yields a family of predicted behaviors

With the terminology in place, we can describe our plan more concretely than before:

- We want to test the hypothesis that the virus evolves within a single patient.
- To this end, we'd like to find the rate at which T cells get infected in the quasi-steady state, because virus mutation is most likely to happen at the error-prone reverse transcription step, which happens once per infection.
- But the production rate of newly infected T cells was not directly measurable in 1995. The measurable quantity was the viral load N_V as a function of time after administering the drug.
- Our model, Equations 1.1–1.2, connects what we've got to what we want, because (*i*) in the quasi-steady state, the infection rate is equal to the rate k_I at which T cells are lost, and (*ii*) k_I is a parameter that we can extract by fitting the model in Equations 1.1–1.2 to data about the *non*-steady state after administering an antiviral drug.

Notice that Equation 1.1 doesn't involve N_V at all; it's one equation in one unknown function, and a very famous one too. Its solution is exponential decay: $N_I(t) = N_{I0}e^{-k_I t}$. The constant N_{I0} is the initial number of infected T cells at time zero. We can just substitute that solution into Equation 1.2 and then forget about Equation 1.1. That is,

$$\frac{dN_V}{dt} = -k_V N_V + \gamma N_{I0} e^{-k_I t}. \tag{1.3}$$

Here k_I, k_V, γ, and N_{I0} are four unknown quantities. But we can simplify our work by noticing that two of them enter the equation only via their product. Hence, we can replace γ and N_{I0} by a single unknown, which we'll abbreviate as $\beta = \gamma N_{I0}$. Because the experiments did not actually measure the population of infected T cells, we need not predict it, and hence we won't need the separate values of γ and N_{I0}.

We could directly solve Equation 1.3 by the methods of calculus, but it's usually best to try for an intuitive understanding first. Think about the metaphor in Figure 0.2, where the volume of water in the middle chamber at any moment plays the role of N_V. For a real leaky container, the rate of outflow depends on the pressure at the bottom, and hence on the level of the water; similarly, Equation 1.2 specifies that the clearance (outflow) rate at time t depends on $N_V(t)$. Next, consider the inflow: Instead of being a constant equal to the outflow, as it is in the steady state, Equation 1.3 gives the inflow rate as $\beta e^{-k_I t}$ (see Figure 0.2).

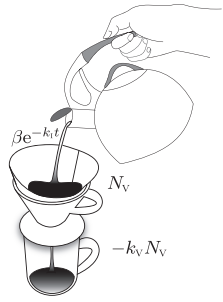


Figure 0.2 (page 2)

Our physical metaphor now lets us guess some general behavior. If $k_I \gg k_V$, then we get a burst of inflow that quickly shuts off, before much has had a chance to run out. So after this brief **transient** behavior, N_V falls exponentially with time, in a way controlled by the decay rate constant k_V. In the opposite extreme case, $k_V \gg k_I$, the container drains nearly as fast as it's being filled;[7] the water level simply tracks the inflow. Thus, again the water level falls exponentially, but this time in a way controlled by the *inflow* decay rate constant k_I.

Our intuition from the preceding paragraph suggests that the long-time behavior of the solution to Equation 1.3 is proportional either to $e^{-k_I t}$ or $e^{-k_V t}$, depending on which rate constant is smaller. We can now try to guess a **trial solution** with this property. In fact, the function

$$N_V(t) \stackrel{?}{=} X e^{-k_I t} + (N_{V0} - X) e^{-k_V t}, \tag{1.4}$$

where X is any constant value, has the desired behavior. Moreover, Equation 1.4 equals N_{V0} at time zero. We now ask if we can choose a value of X that makes Equation 1.4 a solution to Equation 1.3. Substitution shows that indeed it works, if we make the choice $X = \beta/(k_V - k_I)$.

Your Turn 1B

a. Confirm the last statement.
b. The trial solution, Equation 1.4, seems to be the sum of two terms, each of which decreases in time. So how could it have the initial pause that is often seen in data (Figure 0.3)?

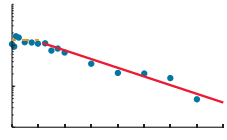


Figure 0.3 (page 4)

T_2 *Section 1.2.4′ (page 21) discusses the hypothesis of viral evolution within a single patient in greater detail.*

1.2.5 Most models must be fitted to data

What has been accomplished so far? We proposed a physical model with three unknown parameters, k_I, k_V, and β. One of these, k_I, is relevant to our hypothesis that virus is rapidly infecting T cells, so we'd like to know its numerical value. Although the population of infected T cells was not directly measurable in 1995, we found that the model makes a

[7] The water never drains completely, because in our model the rate of outflow goes to zero as the height goes to zero. This may not be a good description of a real bucket, but it's reasonable for a virus, which becomes harder for the immune system to find as its concentration decreases.

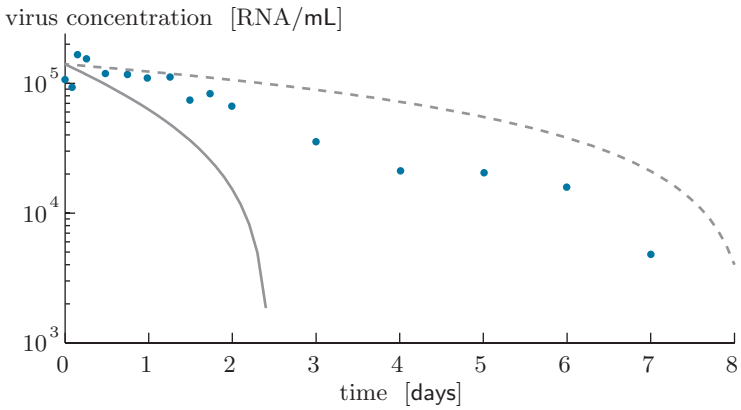


Figure 1.3 [Experimental data.] **Bad fits.** *Blue dots* are the same experimental data that appeared in Figure 0.3 (page 4). The *solid curve* shows the trial solution to our model (Equation 1.4), with a bad set of parameter values. Although the solution starts out at the observed value, it quickly deviates from the data. However, a different choice of parameters does lead to a function that works (see Problem 1.4). The *dashed curve* shows a fit to a different functional form, one not based on any physical model. Although it starts and ends at the right values, in fact, *no* choice of parameters for this model can fit the data.

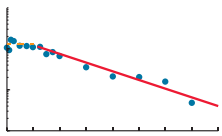


Figure 0.3 (page 4)

prediction for the virus number $N_V(t)$, which *was* observable then. Fitting the model to experimentally measured virus concentration data can thus help us determine the desired rate constant k_I.

The math gave a prediction for $N_V(t)$ in terms of the initial viral load N_{V0} and the values of k_I, k_V, and β. The prediction does have the observed qualitative behavior that after an initial transient, $N_V(t)$ falls exponentially, as seen in Figure 0.3. The remaining challenges are that

- We haven't yet determined the unknowns k_I, k_V, and β.
- The model itself needs to be evaluated critically; all of the assumptions and approximations that went into it are, in principle, suspect.

To gain some confidence in the model, and find the unknown parameter values, we must attempt a detailed comparison with data. We would especially hope to find that different patients, despite having widely different initial viral load N_{V0}, nevertheless all have similar values of k_I. Then the claim that this rate is "very large" (and hence may allow for virus evolution in a single patient) may have some general validity.

Certainly it's not difficult to find things that *don't* work! Figure 1.3 shows the experimental data, along with two functions that don't fit them. One of these functions belongs to the class of trial solutions that we constructed in the preceding section. It looks terrible, but in fact you'll show in Problem 1.4 that a function of this form can be made to look like the data, with appropriate choices of the fitting parameters. The other function shown in the figure is an attempt to fit the data with a simple linear function, $N_V(t) \stackrel{?}{=} A - Bt$. We can make this function pass through our data's starting and ending points, but there is no way to make it fit all of the data. It's not surprising—we have no physical model leading us to expect a linear falloff with time. If we *were* considering such a model, however, our inability to fit it would have to be considered as strong evidence that it is wrong.

After you work through Problem 1.4, you'll have a graph of the best-fitting version of the model proposed in the preceding section. Examining it will let you evaluate the main hypothesis proposed there, by drawing a conclusion from the fit parameter values.

1.2.6 Overconstraint versus overfitting

Our physical model includes crude representations of some processes that we knew must be present in our system, although the model neglects others. It's not perfect, but the agreement between model and data that you'll find in Problem 1.4 is detailed enough to make the model seem "promising." Fitting the model requires that we adjust three unknown parameters, however. There is always the possibility that a model is fundamentally wrong (omits some important features of reality) but nevertheless can be made to *look* good by tweaking the values of its parameters. It's a serious concern, because in that case, the parameter values that gave the fortuitously good fit can be meaningless. Concerns like these must be addressed any time we attempt to model any system.

In our case, however, Problem 1.4 shows that you can fit *more than three* data points by an appropriate choice of three parameters. We say that the data **overconstrain** the model, because there are more conditions to be met than there are parameters to adjust. When a model can match data despite being overconstrained, that fact is unlikely to be a coincidence. The opposite situation is often called **overfitting**; in extreme cases, a model may have so many free fit parameters that it can fit almost any data, regardless of whether it is correct.

Successfully fitting an overconstrained model increases our confidence that it reflects reality, even if it proposes the existence of *hidden actors*, for which we may have little or no direct evidence. In the HIV example, these actors were the T cells, whose population was not directly observable in the original experiments.

T_2 *Section 1.2.6′ (page 21) discusses in more detail the sense in which our model is overdetermined, and outlines a more realistic model for virus dynamics.*

1.3 Just a Few Words About Modeling

Sometimes we examine experimental data, and their form immediately suggests some simple mathematical function. We can write down that function with some parameters, plot it alongside the data, and adjust the parameters to optimize the fit. In Problem 1.5 you'll use this approach, which is sometimes called "blind fitting." Superficially, it resembles what we did with HIV data in this chapter, but there is a key difference.

Blind fitting is often a convenient way to *summarize* existing data. Because many systems respond continuously as time progresses or a parameter is changed, choosing a simple smooth function to summarize data also lets us **interpolate** (predict what would have been measured at points lying between actual measurements). But, as you'll see in Problem 1.5, blind fitting often fails spectacularly at **extrapolation** (predicting what would have been measured at points lying *outside* the range of actual measurements).[8] That's because the mathematical function that we choose may not have any connection to any underlying mechanism giving rise to the behavior.

This chapter has followed a very different procedure. We first imagined a plausible mechanism consistent with other things we knew about the world (a physical model), and then embodied it in mathematical formulas. The physical model may be wrong; for example, it may neglect some important players or interactions. But if it gives nontrivial successful predictions about data, then we are encouraged to test it outside the range of conditions studied in the initial experiments. If it passes those tests as well, then it's "promising,"

[8] Even interpolation can fail: We may have so few data points that a simple function seems to fit them, even though no such relation exists in reality. A third pitfall with blind fitting is that, in some cases, a system's behavior does *not* change smoothly as parameters are changed. Such "bifurcation" phenomena are discussed in Chapter 10.

and we are justified in trying to apply its results to other situations, different from the first experiments. Thus, successful fitting of a model to HIV data suggested a successful treatment strategy (the multidrug treatment described in the Prolog). Other chapters in this book will look at different stories.

In earlier times, a theoretical "model" often just meant some words, or a cartoon. Why must we clutter such images with math? One reason is that equations force a model to be precise, complete, and self-consistent, and they allow its full implications to be worked out, including possible experimental tests. Some "word models" sound reasonable but, when expressed precisely, turn out to be self-inconsistent, or to depend on physically impossible values for some parameters. In this way, modeling can formulate, explore, and often reject potential mechanisms, letting us focus only on experiments that test the promising ones.

Finally, skillful modeling can tell us in advance whether an experiment is likely to be able to discriminate two or more of the mechanisms under consideration, or point to what changes in the experimental design will enhance that ability. For that reason, modeling can also help us make more efficient use of the time and money needed to perform experiments.

THE BIG PICTURE

This chapter has explored a minimal, reductionist approach to a complex biological system. Such an approach has strengths and weaknesses. One strength is generality: The same sort of equations that are useful in understanding the progression of HIV have proven to be useful in understanding other infections, such as hepatitis B and C.

More broadly, we may characterize a good scientific experience as one in which a puzzling result is explained in quantitative detail by a simplified model, perhaps involving some sort of equations. But a *great* scientific experience can arise when you find that some totally different-seeming system or process obeys the *same* equations. Then you gain insights about one system from your experience with the other. In this chapter, the key equations had associations with systems like leaky containers, which helped lead us to the desired solution.

Our physical model of HIV dynamics still has a big limitation, however. Although the fit to data supports the picture of a high virus production rate, this does not completely validate the overall picture proposed in the Prolog (page 4). That proposal went on to assert that *high production somehow leads to the evolution of drug resistance.* In fact, although this connection is intuitive, the framework of this chapter cannot establish it quantitatively. When writing down Equations 1.1–1.2, we tacitly made the assumption that T cell and virus populations were quantities that changed continuously in time. This assumption allowed us to apply some familiar techniques from calculus. But really, those populations are *integers,* and so must change discontinuously in time.

In many situations, the difference is immaterial. Populations are generally huge numbers, and their graininess is too small to worry about. But in our problem we are interested in the rare, chance mutation of *just one* virion from susceptible to resistant (Figure 1.4). We will need to develop some new methods, and intuition, to handle such problems.

The word "chance" in the preceding paragraph highlights another gap in our understanding so far: Our equations, and calculus in general, describe *deterministic* systems, ones for which the future follows inevitably once the present is sufficiently well known. It's an approach that works well for some phenomena, like predicting eclipses of the Sun. But clockwork determinism is not very reminiscent of Life. And even many purely physical phenomena, we will see, are inherently probabilistic in character. Chapters 3–7 will develop the ideas we need to introduce randomness into our physical models of living systems.

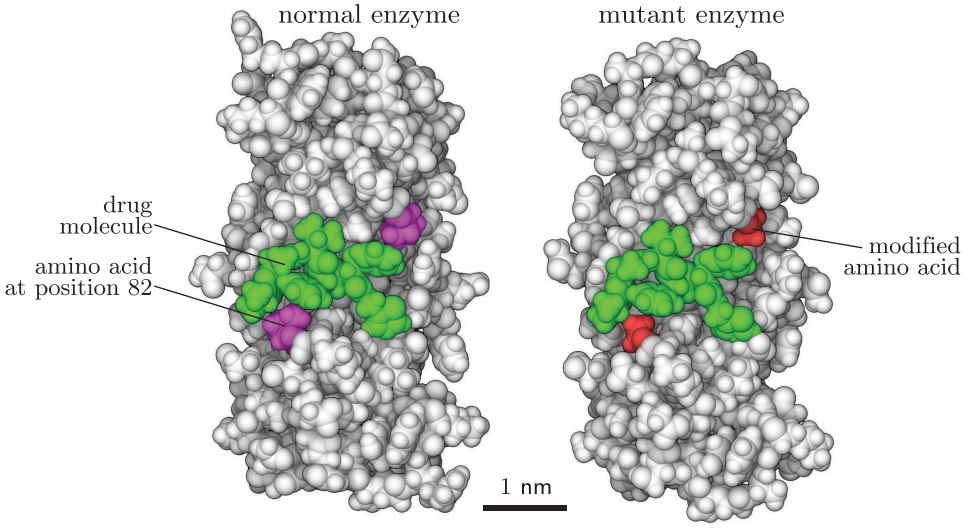


Figure 1.4 [Artist's reconstructions based on structural data.] **Antiviral drugs are molecules that bind tightly to HIV's enzymes** and block their action. HIV protease can become resistant to drugs by mutating certain of its amino acids; such a mutation changes its shape slightly, degrading the fit of the drug molecule to its usual binding site. The drug ritonavir is shown in *green. Left:* The amino acids at position 82 in each of the enzyme's two protein chains are normally valines (*magenta*); they form close contacts with the drug, stabilizing the binding. The bound drug molecule then obstructs the enzyme's active site, preventing it from carrying out its function (see Figure 1.1). *Right:* In the mutant enzyme, this amino acid has been changed to the smaller alanine (*red*), weakening the contact slightly. Ritonavir then binds poorly, and so does not interfere with the enzyme's activity even when it is present. [Courtesy David S Goodsell.]

KEY FORMULAS

Throughout the book, closing sections like this one will collect useful formulas that appeared in each chapter. In this chapter, however, the section also includes formulas from your previous study of math that will be needed later on.

- *Mathematical results:* Make sure you recall these formulas and how they follow from Taylor's theorem. Some are valid only when x is "small" in some sense.

$$\exp(x) = 1 + x + \cdots + \frac{1}{n!}x^n + \cdots$$

$$\cos(x) = 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \cdots$$

$$\sin(x) = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \cdots$$

$$1/(1-x) = 1 + x + \cdots + x^n + \cdots$$

$$\ln(1-x) = -x - \frac{1}{2}x^2 - \cdots - \frac{1}{n}x^n - \cdots$$

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \cdots$$

In addition, we'll later need these formulas:
The binomial theorem: $(x+y)^M = C_{M,0}x^M y^0 + C_{M,1}x^{M-1}y^1 + \cdots + C_{M,M}x^0 y^M$, where the binomial coefficients are given by

$$C_{M,\ell} = M!/\big(\ell!(M-\ell)!\big) \text{ for } \ell = 0, \ldots, M.$$

The Gaussian integral: $\int_{-\infty}^{\infty} \mathrm{d}x \, \exp(-x^2) = \sqrt{\pi}$.
The compound interest formula:[9] $\lim_{M\to\infty} \left(1 + \frac{a}{M}\right)^M = \exp(a)$.

- *Continuous growth/decay:* The differential equation $\mathrm{d}N_\mathrm{I}/\mathrm{d}t = kN_\mathrm{I}$ has solution $N_\mathrm{I}(t) = N_{\mathrm{I}0}\exp(kt)$, which displays exponential decay (if k is negative) or growth (if k is positive).
- *Viral dynamics model:* After a patient begins taking antiviral drugs, we proposed a model in which the viral load and population of infected T cells are solutions to

$$\frac{\mathrm{d}N_\mathrm{I}}{\mathrm{d}t} = -k_\mathrm{I}N_\mathrm{I} \qquad \text{for } t \geq 0, \tag{1.1}$$

$$\frac{\mathrm{d}N_\mathrm{V}}{\mathrm{d}t} = -k_\mathrm{V}N_\mathrm{V} + \gamma N_\mathrm{I}. \tag{1.2}$$

FURTHER READING

Semipopular:
On overfitting: Silver, 2012, chapt. 5.

Intermediate:
HIV: Freeman & Herron, 2007.
Modeling and HIV dynamics: Ellner & Guckenheimer, 2006, §6.6; Nowak, 2006; Otto & Day, 2007, chapt. 1; Shonkwiler & Herod, 2009, chapt. 10.

Technical:
Ho et al., 1995; Nowak & May, 2000; Perelson & Nelson, 1999; Wei et al., 1995. Equations 1.1 and 1.2 appeared in Wei et al., 1995, along with an analysis equivalent to Section 1.2.4.

[9] The left side of this formula is the factor multiplying an initial balance on a savings account after one year, if interest is compounded M times a year at an annual interest rate a.

T2 **Track 2**

1.2.4′ Exit from the latency period

Figure 0.1 (page 1)

Prior to the events of 1995, Nowak, May, and Anderson had already developed a theory for the general behavior shown in Figure 0.1 (see Nowak, 2006, chapt. 10). According to this theory, during the initial spike in viral load, one particular strain of HIV becomes dominant, because it reproduces faster than the others. The immune system manages to control this one strain, but over time it mutates, generating diversity. Eventually, the immune system gets pushed to a point beyond which it is unable to cope simultaneously with all the strains that have evolved by mutation, and the virus concentration rises rapidly. Meanwhile, each round of mutation stimulates a new class of T cells to respond, more and more of which are already infected, weakening their response.

T2 **Track 2**

1.2.6′a Informal criterion for a falsifiable prediction

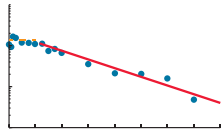


Figure 0.3 (page 4)

The main text stated that our result was significant because we could fit many (more than three) data points by adjusting only three unknown parameters: k_I, k_V, and β. It's a bit more precise to say that the data in Figure 0.3 have several independent "visual features": the slope and intercept of the final exponential decay line, the initial slope and value N_{V0}, and the sharpness of the transition from initial plateau to exponential decay. Of these five features, N_{V0} was already used in writing the solution, leaving four that must be fit by parameter choice. But we have only three parameters to adjust, which in principle makes our trial solution a **falsifiable prediction**: There is no mathematical guarantee that *any* choice of parameters can be found that will fit such data. If we do find a set of values that fit, we may at least say that the data have missed an opportunity to falsify the model, increasing our confidence that the model may be correct. In this particular case, none of the visual features are very precisely known, due to scatter in the data and the small number of data points available. Thus, we can only say that (*i*) the data are not qualitatively inconsistent with the model, but (*ii*) the data *are* inconsistent with the value $(k_I)^{-1} \approx 10$ years suggested by the hypothesis of a slow virus.

1.2.6′b More realistic viral dynamics models

Many improvements to the model of HIV dynamics in the main text have been explored. For example, we supposed that no new infections occur after administering the drug, but neither of the drug classes mentioned in the text work in exactly this way. Some instead block reverse transcription after virus entry; such a drug may be only partially effective, so that new infections of T cells continue, at a reduced rate, after administration of the drug. Other drugs seek to stop the production of "competent" virions; these, too, may be only partly effective. A more complex set of equations incorporating these ideas appeared in Perelson (2002). Letting $N_U(t)$ be the population of uninfected T cells and $N_X(t)$ that of inactive virions, the model becomes

$$\frac{dN_U}{dt} = \lambda - k_V N_U - \epsilon N_V N_U, \tag{1.5}$$

$$\frac{dN_I}{dt} = \epsilon N_V N_U - k_I N_I, \tag{1.6}$$

$$\frac{dN_V}{dt} = \epsilon' \gamma N_I - k_V N_V, \tag{1.7}$$

$$\frac{dN_X}{dt} = (1 - \epsilon')\gamma N_I - k_V N_X. \tag{1.8}$$

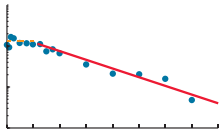


Figure 0.3 (page 4)

Here, the normal birth and death of T cells are described by the constants λ and k_V, respectively, the residual infectivity by ϵ, and the fraction of competent virions produced by ϵ'.

Even this more elaborate model makes assumptions that are not easy to justify. But it can account nicely for a lot of data, including the fact that at longer times than those shown in Figure 0.3, virus concentration stops dropping exponentially.

1.2.6′c Eradication of HIV

Unfortunately, not all infected cells are short lived. The cell death rate that we found from the math reflects only the subset of infected cells that actively add virions to the blood; but some infected cells do not. Some of the other, nonproductive infected cells have a latent "provirus" in their genome, which can be activated later. That's one reason why complete eradication of HIV remains difficult.

PROBLEMS

1.1 Molecular graphics

You can make your own molecular graphics. Access the Protein Data Bank (Media 3[10]). The simplest way to use it is to enter the name or code of a macromolecule in the search box. Then click "3D view" on the right and manipulate the resulting image. Alternatively, you can download coordinate data for the molecule to your own computer and then visualize it by using one of the many free software packages listed in Media 3. To find some interesting examples, you can explore the past Molecules of the Month or see page 321 for the names of entries used in creating images in this book.

a. Make images based on the following entries, which are relevant to the present chapter:
   - `1jlb` (HIV-1 reverse transcriptase in complex with nevirapine)
   - `1hsg` (HIV-2 protease complexed with a protease inhibitor)
   - `1rl8` (resistant strain of HIV-1 protease with ritonavir)

b. Now try these entries, which are molecules to be discussed in Chapter 10:
   - `1lbh` (*lac* repressor bound to the gratuitous inducer IPTG)
   - `3cro` (Cro transcription factor, bound to a segment of DNA)
   - `1pv7` (lactose permease bound to the lactose-like molecule TDG)

c. These entries are also interesting:
   - `1mme` (hammerhead ribozyme)
   - `2f8s` (small interfering RNA)

1.2 Semilog and log-log plots

a. Use a computer to plot the functions $f_1(x) = \exp(x)$ and $f_2(x) = x^{3.5}$ on the range $2 \le x \le 7$. These functions may appear qualitatively similar.

b. Now make semilogarithmic graphs of the same two functions. What outstanding feature of the exponential function jumps out in this representation?

c. Finally, make log-log graphs of the two functions and comment.

1.3 Half-life

Sometimes instead of quoting a rate constant like k_I in Equation 1.1 (page 14), scientists will quote a **half-life**, the time after which an exponentially falling population has decreased to half its original value. Derive the relation between these two quantities.

1.4 Model action of antiviral drug

Finish the analysis of the time course of HIV infection after administering an antiviral drug. For this problem, you may assume that virus clearance is faster than T cell death (though not necessarily much faster). That is, assume $k_V > k_I$.

a. Follow Section 1.2.4 (page 14) to write down the trial solution for $N_V(t)$, the observable quantity, in terms of the initial viral load N_{V0} and three unknown constants k_I, k_V, and β.

b. Obtain Dataset 1,[11] and use a computer to make a semilog plot. Don't join the points by line segments; make each point a symbol, for example, a small circle or plus sign. Label the axes of your plot. Give it a title, too. Superimpose a graph of the trial solution, with

[10] References of this form refer to Media links on the book's Web site.
[11] References of this form refer to Dataset links on the book's Web site.

some arbitrary values of k_I, k_V, and β, on your graph of the actual data. Then try to fit the trial solution to the data, by choosing better values of the parameters.

c. You may quickly discover that it's difficult to find the right parameter values just by guessing. Rather than resort to some black-box software to perform the search, however, try to choose parameter values that make certain features of your graph coincide with the data, as follows. First, note that the experimental data approach a straight line on a semilog plot at long times (apart from some experimental scatter). The trial solution Equation 1.4 also approaches such a line, namely, the graph of the function $N_{V,asymp}(t) = Xe^{-k_I t}$, so you can match that function to the data. Lay a ruler along the data, adjust it until it seems to match the long-time trend of the data, and find two points on that straight line. From this information, find values of k_I and X that make $N_{V,asymp}(t)$ match the data.

d. Substitute your values of k_I and X into Equation 1.4 (page 15) to get a trial solution with the right initial value N_{V0} and the right long-time behavior. This is still not quite enough to give you the value of k_V needed to specify a unique solution. However, the model suggests another constraint. Immediately after administering the drug, the number of infected T cells has not yet had a chance to begin falling. Thus, in this model both viral production and clearance are the same as they were prior to time zero, so the solution is initially still quasi-steady:

$$\left.\frac{dN_V}{dt}\right|_{t=0} = 0.$$

Use this constraint to determine all parameters of the trial solution from your answer to (c), and plot the result along with the data. (You may want to tweak the approximate values you used for N_{V0}, and other parameters, in order to make the fit look better.)

e. The hypothesis that we have been exploring is that the reciprocal of the T cell infection rate is much shorter than the typical latency period for the infection, or in other words that

$$(1/k_I) \text{ is much smaller than } 10 \text{ years}.$$

Do the data support this claim?

f. Use your result from Problem 1.3 to convert your answers to (d) into half-life values for virions and infected T cells. These represent half-lives in a hypothetical system with clearance, but no new virion production nor new infections.

1.5 Blind fitting

Obtain Dataset 2. This file contains an array consisting of two columns of data. The first is the date in years after an arbitrary starting point. The second is the estimated world population on that date.

a. Use a computer to plot the data points.

b. Here is a simple mathematical function that roughly reproduces the data:

$$f(t) = \frac{100\,000}{2050 - (t/1\,\text{year})}. \tag{1.9}$$

Have your computer draw this function, and superimpose it on your graph of the actual data. Now play around with the function, trying others of the same form

$f(t) = A/(B - t)$, for some constants A and B. You can get a pretty good-looking fit in this way. (There are automated ways to do this, but it's instructive to try it "by hand" at least once.)

c. Do you think this is a good model for the data? That is, does it tell us anything interesting beyond roughly reproducing the data points? Explain.

1.6 T2 Special case of a differential equation system

Consider the following system of two coupled linear differential equations, simplified a bit from Equations 1.1 and 1.2 on page 14:

$$dA/dt = -k_A A \qquad \text{and} \qquad dB/dt = -k_B B + A.$$

This set of equations has two linearly independent solutions, which can be added in any linear combination to get the general solution.[12] So the general solution has two free parameters, the respective amounts of each independent solution.

Usually, a system of linear differential equations with constant coefficients has solutions in the form of exponential functions. However, there is an exceptional case, which can arise even in the simplest example of two equations. Remember the physical analogy for this problem: The first equation determines a function $A(t)$, which is the flow rate into a container B, which in turn has a hole at the bottom.

Section 1.2.4 argued that if $k_A \ll k_B$, then B can't accumulate much, and its outflow is eventually determined by k_A. In the opposite case, $k_B \ll k_A$, B fills up until A runs out, and then trickles out its contents in a way controlled by k_B. Either way, the long-time behavior is an exponential, and in fact we found that at *all* times the behavior is a combination of two exponentials.

But what if $k_A = k_B$? The above reasoning is inapplicable in that case, so we can't be sure that every solution falls as an exponential at long times. In fact, there is *one* exponential solution, which corresponds to the situation where $A(0) = 0$, so we have only B running out. But there must also be a second, independent solution.

a. Find the other solution when $k_A = k_B$, and hence find the complete general solution to the system. Get an analytic result (a formula), not a numerical one. [*Hint:* Solve container A's behavior explicitly: $A(t) = A_0 \exp(-k_A t)$, and substitute into the other equation to get

$$dB/dt = -k_A B + A_0 \exp(-k_A t).$$

If $A_0 \neq 0$, then no solution of the form $B(t) = \exp(-k_A t)$ works. Instead, play around with multiplying that solution by various powers of t until you find something that solves the equation.]

b. Why do you suppose this case is *not* likely to be relevant to a real-life problem like our HIV story?

1.7 T2 Infected cell count

First, work Problem 1.4. Then continue as follows to get an estimate of the population of infected T cells in the quasi-steady state. Chapter 3 will argue that this number is needed in order to evaluate the hypothesis of viral evolution in individual patients.

[12] In general, a system of N first-order linear differential equations in N unknowns has N independent solutions.

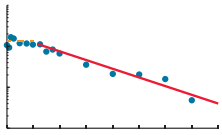
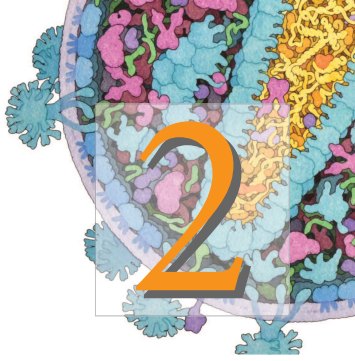


Figure 0.3 (page 4)

a. The human body contains about 5 L of blood. Each HIV virion carries two copies of its RNA genome. Thus, the total virion population is about $2.5 \cdot 10^3$ mL times the quantity plotted in Figure 0.3. Express the values of N_{V0} and X that you found in Problem 1.4 in terms of total virion population.

b. Obtain a numerical estimate for β from your fit. (You found the value of k_I in Problem 1.4.)

c. The symbol β is an abbreviation for the product γN_{I0}, where $\gamma \approx 100k_I$ is the rate of virion release by an infected T cell and N_{I0} is the quantity we want to find. Turn your results from (a) and (b) into an estimate of N_{I0}.



2

Physics and Biology

It is not the strongest of the species that survives, nor the most intelligent, but rather the one most responsive to change.
—Charles Darwin

2.1 Signpost

The rest of this book will explore two broad classes of propositions:

1a. Living organisms use physical mechanisms to gain information about their surroundings, and to respond to that information. Appreciating the basic science underlying those mechanisms is critical to understanding how they work.

1b. Scientists also use physical mechanisms to gain information about the systems they are studying. Here, too, appreciating some basic science allows us to extend the range and validity of our measurements (and, in turn, of the models that those measurements support).

2a. Living organisms must make inferences (educated guesses) about the best response to make, because their information is partly diluted by noise.[1]

2b. In many cases scientists, too, must reason probabilistically to extract the meaning from our measurements.

In fact, the single most characteristic feature of living organisms, cutting across their immense diversity, is their adaptive response to the opportunities and constraints of their ever-fluctuating physical environment. Organisms must gather *information* about the world, make *inferences* about its present and future states based on that information, and *modify behavior* in ways that optimize some outcome.

[1] The everyday definition of **noise** is "uninformative audio stimuli," or more specifically "music enjoyed by people in any generation other than my own." But in this book, the word is a synonym for "randomness," defined in Chapter 3.

Each of the propositions above has been written in two parallel forms, in order to highlight a nice symmetry:

> *The same sort of probabilistic inference needed to understand your lab data must also be used by the lion and the gazelle as they integrate their own data and make their decisions.*

2.2 The Intersection

At first sight, Physics may seem to be almost at the opposite intellectual pole from Biology. On one side, we have Newton's three simple laws;[2] on the other, the seemingly arbitrary Tree of Life. On one side, there is the relentless search for simplicity; on the other, the appearance of irreducible complexity. One side's narrative stresses universality; the other's seems dominated by historical accidents. One side stresses determinism, the prediction of the future from measurement of the present situation; the other's reality is highly unpredictable.

But there is more to Physics than Newton's laws. Gradually during the 19th century, scientists came to accept the lumpy (molecular) character of all matter. At about the same time, they realized that if the air in a room consists of tiny particles independently flying about, that motion must be *random*—not deterministic. We can't see this motion directly, but by the early 20th century it became clear that it was the cause of the incessant, random motion of any micrometer-size particle in water (called **Brownian motion**; see Media 2). A branch of Physics accordingly arose to describe such purely physical, yet random, systems. It turned out that conclusions can be drawn from intrinsically random behavior, essentially because every sort of "randomness" actually has characteristics that we can measure quantitatively, and even try to predict. Similar methods apply to the randomness found in living systems.

Another major discovery of early 20th century Physics was that light, too, has a lumpy character. Just as we don't perceive that a stream of water consists of discrete molecules, so too in normal circumstances we don't notice the granular character of a beam of light. Nevertheless, that aspect will prove essential in our later study of localization microscopy.

Turning now to Biology, the great advance of the late 19th century was the principle of *common descent:* Because all living organisms partially share their family tree, we can learn about any of them by studying any other one. Just as physicists can study simple atoms and hope to find clues about the most complex molecule, so too could biologists study bacteria with reasonable expectations of learning clues about butterflies and giraffes. Moreover, inheritance along that vast family tree has a particulate character: It involves discrete lumps of information (genes), which are either copied exactly or else suffer random, discrete changes (mutation or recombination). This insight was extensively developed in the 20th century; it became clear that, although the long-run *outcomes* of inheritance are subtle and gorgeously varied, many of the underlying *mechanisms* are universal throughout the living world.

Individuals within a species meet, compete, and mate at least partially at random, in ways that may remind us of the physical processes of chemical reactions. Within each individual, too, each cell's life processes are literally chemical reactions, again involving discrete molecules. Some of the key actors in this inner dance appear in only a very few copies. Some respond to external signals that involve only a few discrete entities. For example,

[2] And Maxwell's less simple, but still concise, equations.

olfactory (smell) receptors can respond to just a few odorant molecules; visual receptors can respond to the absorption of even *one* unit of light. At this deep level, the distinction between biological and physical science melts away.

In short, 20th century scientists gained an ever increasing appreciation of the fact that

Discreteness and randomness lie at the roots of many physical and biological phenomena.

They developed mathematical techniques appropriate to both realms, continuing to the present.

2.3 Dimensional Analysis

Appendix B describes an indispensable tool for organizing our thoughts about physical models. Here are two quick exercises in this area that are relevant to topics in this book.

Your Turn 2A

Go back to Equation 1.3 (page 14) and check that it conforms to the rules for units given in Appendix B. Along the way, find the appropriate units for the quantities k_I, k_V, β, and γ. From that, show that statements like "k_V is much larger than $1/(10\,\text{years})$" indeed make sense dimensionally.

Your Turn 2B

Find the angular diameter of a coin held at a distance of 3 m from your eye. (Take the coin's diameter to be 1 cm.) Express your answer in radians and in arc minutes. Compare your answer to the angular diameter of the Moon when viewed from Earth, which is about 32 arcmin.

In this book, the names of units are set in a special typeface, to help you distinguish them from named quantities. Thus cm denotes "centimeters," whereas *cm* could denote the product of a concentration times a mass, and "cm" could be an abbreviation for some ordinary word. Symbols like $\mathbb{L}$ denote dimensions (in this case, length); see Appendix B.

Named quantities are generally single italicized letters. We can assign them arbitrarily, but we must use them consistently, so that others know what we mean. Appendix A collects definitions of many of the named quantities, and other symbols, used in this book.

THE BIG PICTURE

In physics classes, "error analysis" is sometimes presented as a distasteful chore needed to overcome some tiresome professor's (or peer reviewer's) objections to our work. In the biology curriculum, it's sometimes relegated to a separate course on the design of clinical trials. This book will instead try to integrate probabilistic reasoning directly into the study of how living organisms manage their amazing trick of responding to their environment.

Looking through this lens, we will study some case histories of responses at many levels and on many time scales. As mentioned in the Prolog, even the very most primitive life forms (viruses) respond at the population level by evolving responses to novel challenges,

and so do all higher organisms. Moving up a step, individual bacteria have genetic and metabolic circuits that endow them with faster responses to change, enabling them to turn on certain capabilities only when they are needed, become more adaptable in hard times, and even search for food. Much more elaborate still, we vertebrates have exceedingly fast neural circuits that let us hit a speeding tennis ball, or snag an insect with our long, sticky tongue (as appropriate). Every level involves physical ideas. Some of those ideas may be new to you; some seem to fly in the face of common sense. (You may need to change and adapt a bit yourself to get a working understanding of them.)

KEY FORMULAS

See also Appendix B.

- *Angles:* To find the angle between two rays that intersect at their end points, draw a circular arc, centered on the common end point, that starts on one ray and ends on the other one. The angle in radians (rad) is the ratio of the length of that arc to its radius. Thus, angles, and the unit rad, are dimensionless. Another dimensionless unit of angle is the degree, defined as $\pi/180$ radians.

FURTHER READING

Here are four books that view living organisms as information-processing machines:

Semipopular:
Bray, 2009.

Intermediate:
Alon, 2006; Laughlin & Sterling, 2015.

Technical:
Bialek, 2012.

PROBLEMS

2.1 Greek to me

We'll be using a lot of letters from the Greek alphabet. Here are the letters most often used by scientists. The following list gives both lowercase and uppercase (but omits the uppercase when it looks just like a Roman letter):

$$\alpha,\ \beta,\ \gamma/\Gamma,\ \delta/\Delta,\ \epsilon,\ \zeta,\ \eta,\ \theta/\Theta,\ \kappa,\ \lambda/\Lambda,\ \mu,\ \nu,$$

$$\xi/\Xi,\ \pi/\Pi,\ \rho,\ \sigma/\Sigma,\ \tau,\ \upsilon/\Upsilon,\ \phi/\Phi,\ \chi,\ \psi/\Psi, \text{and } \omega/\Omega$$

When writing computer code, we often spell them out as alpha, beta, gamma, delta, epsilon, zeta, eta, theta, kappa, lambda, mu, nu, xi (pronounced "k'see"), pi, rho, sigma, tau, upsilon, phi, chi (pronounced "ky"), psi, and omega, respectively.

Practice by examining the following quote:

> Cell and tissue, shell and bone, leaf and flower, are so many portions of matter, and it is in obedience to the laws of physics that their particles have been moved, moulded, and conformed. They are no exception to the rule that $\Theta\epsilon\grave{o}\varsigma\ \alpha\epsilon\grave{\iota}\ \gamma\epsilon\omega\mu\epsilon\tau\rho\epsilon\hat{\iota}$. – D'Arcy Thompson

From the sounds made by each letter, can you guess what Thompson was trying to say? [*Hint:* ς is an alternate form of σ.]

2.2 Unusual units

In the United States, automobile fuel consumption is usually quantified by stating the car's "miles per gallon" rating. In some ways, the reciprocal of this quantity, called "fuel efficiency," is more meaningful. State the dimensions of fuel efficiency, and propose a natural SI unit with those dimensions. Give a physical/geometrical interpretation of the fuel efficiency of a car that gets 30 miles per gallon of gasoline.

2.3 Quetelet index

It's straightforward to diagnose obesity if a subject's percentage of body fat is known, but this quantity is not easy to measure. Frequently the "body mass index" (BMI, or "Quetelet index") is used instead, as a rough proxy. BMI is defined as

$$\text{BMI} = \frac{\text{body mass in kilograms}}{(\text{height in meters})^2},$$

and BMI > 25 is sometimes taken as a criterion for overweight.

a. Re-express this criterion in terms of the quantity m/h^2. Instead of the pure number 25, your answer will involve a number with dimensions.

b. What's wrong with the simpler criterion that a subject is overweight if body mass m exceeds some fixed threshold?

c. Speculate why the definition above for BMI might be a better, though not perfect, proxy for overweight.

2.4 Mechanical sensitivity

Most people can just barely feel a single grain of salt dropped on their skin from a height $h = 10\,\text{cm}$. Model a grain of salt as a cube of length about 0.2 mm made of a material of

mass density about $10^3\ \mathrm{kg\,m^{-3}}$. How much gravitational potential energy does that grain release when it falls from 10 cm? [*Hint:* If you forget the formula, take the parameters given in the problem and use dimensional analysis to find a quantity with the units $\mathrm{J} = \mathrm{kg\,m^2 s^{-2}}$. Recall that the acceleration due to gravity is $g \approx 10\ \mathrm{m\,s^{-2}}$.]

2.5 Do the wave

a. Find an approximate formula for the speed of a wave on the surface of the deep ocean. Your answer may involve the mass density of water ($\rho \approx 10^3\ \mathrm{kg\,m^{-3}}$), the wavelength λ of the wave, and/or the acceleration of gravity ($g \approx 10\ \mathrm{m\,s^{-2}}$). [*Hints:* Don't work hard; don't write or solve any equation of motion. The depth of the ocean doesn't matter (it's essentially infinite), nor do the surface tension or viscosity of the water (they're negligible).]

b. Evaluate your answer numerically for a wavelength of one meter to see if your result is reasonable.

2.6 Concentration units

Appendix B introduces a unit for concentration called "molar," abbreviated M. To practice dimensional analysis, consider a sugar solution with concentration 1 mM. Find the average number of sugar molecules in one cubic micrometer of such a solution.

2.7 Atomic energy scale

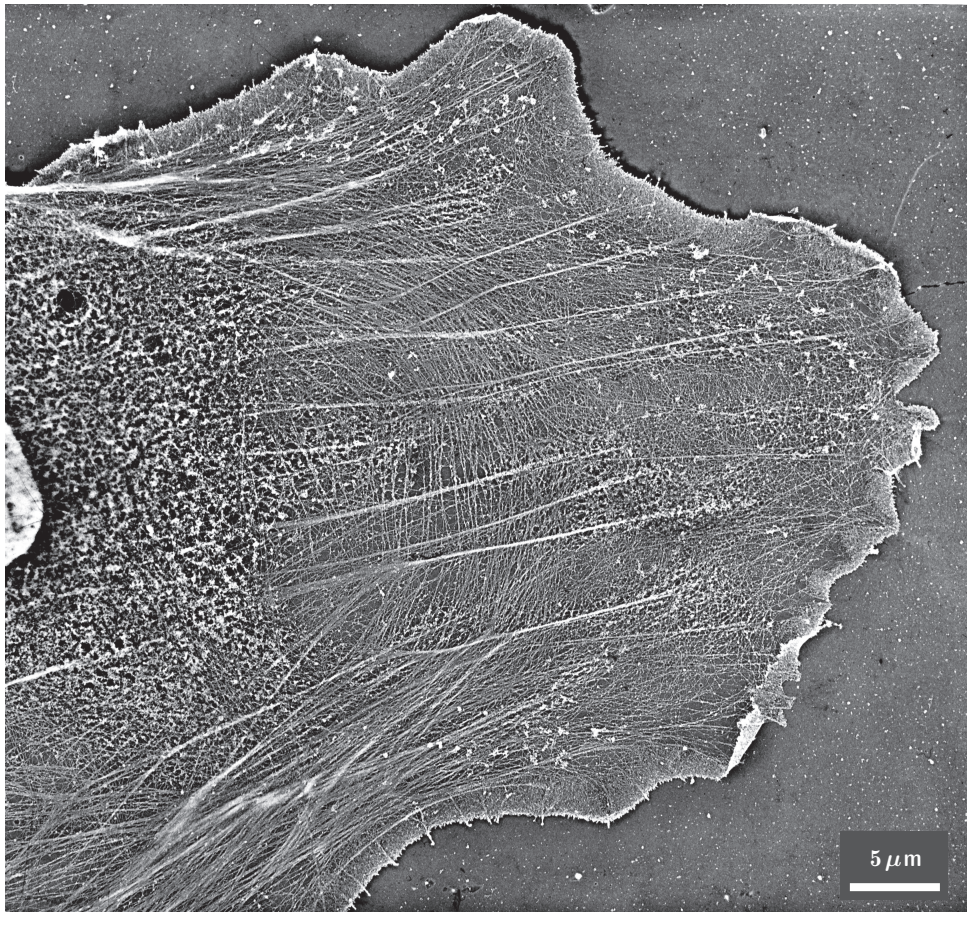
Read Appendix B.

a. Using the same logic as in Section B.6, try to construct an *energy* scale as a combination of the force constant k_e defined there, the electron mass m_e, and Planck's constant $\hbar$. Get a numerical answer in joules. What are the values of the exponents a, b, and c analogous to Equation B.1 (page 314)?

b. We know that chemical reactions involve a certain amount of energy per molecule, which is generally a few eV, where the **electron volt** unit is 1.6×10^{-19} J. (For example, the energy needed to remove the electron from a hydrogen atom is about 14 eV.) How well does your estimate in (a) work?

PART II

Randomness in Biology



[Electron micrograph.] **The leading edge of a crawling cell** (a fibroblast from the frog *Xenopus laevis*). An intricate network of filaments (the cytoskeleton) has been highlighted. Although it is not perfectly regular, neither is this network perfectly random—in any region, the distribution of filament orientations is well defined and related to the cell's function. [Courtesy Tatyana Svitkina, University of Pennsylvania.]

Discrete Randomness

Chance is a more fundamental conception than causality.
—Max Born

3.1 Signpost

Suppose that 30 people are gathered for a meeting. We organize them alphabetically by first name, then list each person's height in that order. It seems intuitive that the resulting list of numbers is "random," in the sense that there is no way to predict any of the numbers. But certainly the list is not "totally random"—we can predict in advance that there will be no heights exceeding, say, 3 m. No series of observations is ever totally unpredictable.

It also makes intuitive sense to say that if we sort the list in ascending order of height, it becomes "less random" than before: Each entry is known to be no smaller than its predecessor. Moreover, if the first 25 heights in the alphabetical list are all under 1 m, then it seems reasonable to draw some tentative conclusions about those people (probably they are children), and even about person number 26 (probably also a child).

This chapter will distill intuitions like these in a mathematical framework general enough for our purposes. This systematic study of randomness will pay off as we begin to construct physical models of living systems, which must cope with (*i*) randomness coming from their external environment and (*ii*) intrinsic randomness from the molecular mechanisms that implement their decisions and actions.

Many of our discussions will take the humble coin toss as a point of departure. This may not seem like a very biological topic. But the coin toss will lead us directly to some less trivial random distributions that do pervade biology, for example, the Binomial, Poisson, and Geometric distributions. It also gives us a familiar setting in which to frame more general ideas, such as likelihood; starting from such a concrete realization will keep our feet on the ground as we generalize to more abstract problems.

This chapter's Focus Question is
Biological question: If each attempt at catching prey is an independent random trial, how long must a predator wait for its supper?
Physical idea: Distributions like this one arise in many physical contexts, for example, in the waiting times between enzyme turnovers.

3.2 Avatars of Randomness

3.2.1 Five iconic examples illustrate the concept of randomness

Let's consider five concrete physical systems that yield results commonly described as "random." Comparing and contrasting the examples will help us to build a visual vocabulary to describe the kinds of randomness arising in Nature:

1. We flip a coin and record which side lands upward (heads or tails).
2. We evaluate the integer random number function defined in a computer math package.
3. We flip a coin m times and use the results to construct a "random, m-bit binary fraction," a number analogous to the familiar decimal fractions:

$$x = \frac{1}{2}s_1 + \frac{1}{4}s_2 + \cdots + \frac{1}{2^m}s_m, \tag{3.1}$$

   where $s_i = 1$ for heads or 0 for tails. x is always a number between 0 and 1.
4. We observe a very dim light source using a sensitive light detector. The detector responds with individual electrical impulses ("blips"), and we record the elapsed waiting time t_w between each blip and its predecessor.[1]
5. We observe the successive positions of a micrometer-size particle undergoing free motion in water (Brownian motion) by taking video frames every few seconds.[2]

Let's look more closely at these examples, in turn. We'd like to extract a general idea of randomness, and also learn how to characterize different *kinds* of randomness.

1a. Actually, coin flipping is not intrinsically unpredictable: We can imagine a precision mechanical coin flipper, isolated from air currents, that reliably results in heads landing up every time. Nevertheless, when a human flips a coin, we do get a series $s_1, s_2, \ldots$ that has no *discernible, relevant* structure: Apart from the constraint that each s_i has only two allowed values, it is essentially unpredictable. Even if we construct an unfair coinlike object that lands heads some fraction ξ of the time, where $\xi \neq 1/2$, nevertheless that one number completely characterizes the resulting series. We will refer often to this kind of randomness, which is called a **Bernoulli trial**. As long as ξ does not equal 1 or 0, we cannot completely predict any result from its predecessors.

2a. A computer-generated series of "random" numbers also cannot literally be random—computers are designed to give perfectly deterministic answers to mathematical calculations. But the computer's algorithm yields a sequence so complex that, for nearly any practical purpose, it too has no discernible, relevant structure.

3a. Turning to the binary fraction example, consider the case of double flips of a fair coin ($m = 2$ and $\xi = 1/2$). There are four possible outcomes, namely, *TT*, *TH*, *HT*, and

[1] You can listen to a sample of these blips (Media 5) and contrast it with a sample of regularly spaced clicks with the same average rate (Media 6).
[2] See Media 2.

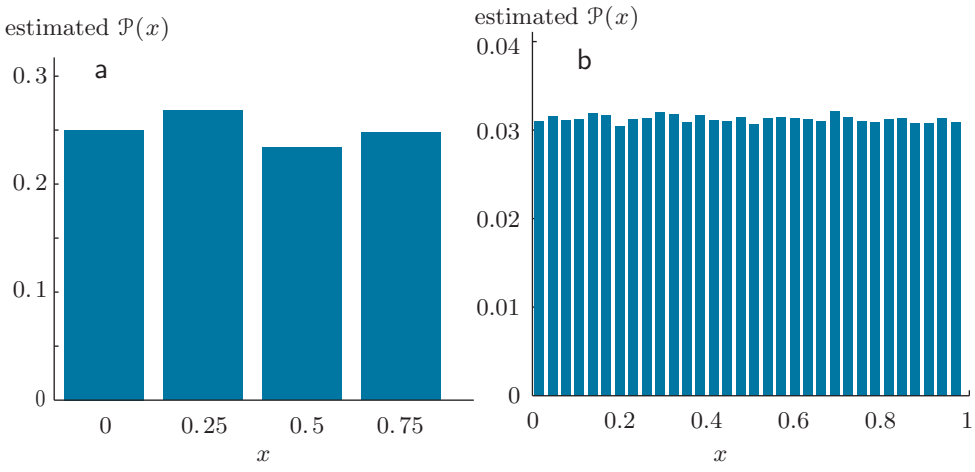


Figure 3.1 [Computer simulations.] **Uniformly distributed random variables.** Empirical distributions of (a) 500 two-bit random binary fractions (that is, $m = 2$ in Equation 3.1) and (b) 250 000 five-bit binary fractions ($m = 5$). The symbol $\mathcal{P}(x)$ refers to the probabilities of various outcomes; it will be defined precisely later in this chapter.

HH, which yield the numbers $x = 0$, 1/4, 1/2, and 3/4, respectively. If we make a lot of double flips and draw a histogram of the results (Figure 3.1a), we expect to find four bars of roughly equal height: The successive x values are drawn from a **discrete Uniform distribution**. If we choose a larger value of m, say, $m = 5$, we find many more possible outcomes; the allowed x values are squeezed close to one another, staying always in the range from $x = 0$ to 1 (Figure 3.1b). Again, the bars in the resulting histogram are all roughly equal in height (if we have measured a large enough number of instances).[3] All we can say about the next number drawn from this procedure is that $0 \leq x < 1$, but even that is *some* knowledge.

4a. For the light detector example, no matter how hard we work to improve our apparatus, we always get irregularly spaced blips at low light intensity. However, we do observe a certain amount of structure in the intervals between successive light detector blips: These waiting times t_w are always greater than zero, and, moreover, short waits are more common than long ones (Figure 3.2a). That is, the thing that's predictable is, again, a *distribution*—in this case, of the waiting times. The "cloud" representation shown in the figure makes it hard to say anything more precise than that the outcomes are not all equally probable. But a histogram-like representation (Figure 3.2b) reveals a definite form. Unlike example **3** above, the figures show that this time the distribution is *non-Uniform:* It's an example of an "Exponential distribution."[4]

So in this case, we again have limited knowledge, obtained from our experience with many previous experiments, that helps us to guess the waiting time before the next blip. We can learn a bit more by examining, say, the first 100 entries in the series of waiting

[3] We can even imagine a limit of very large m; now the tops of the histogram bars nearly form a continuous line, and we are essentially generating a sequence of real numbers drawn from the **continuous Uniform distribution** on the interval. Chapter 5 will discuss continuous distributions.

[4] Chapter 7 will discuss this distribution in detail.

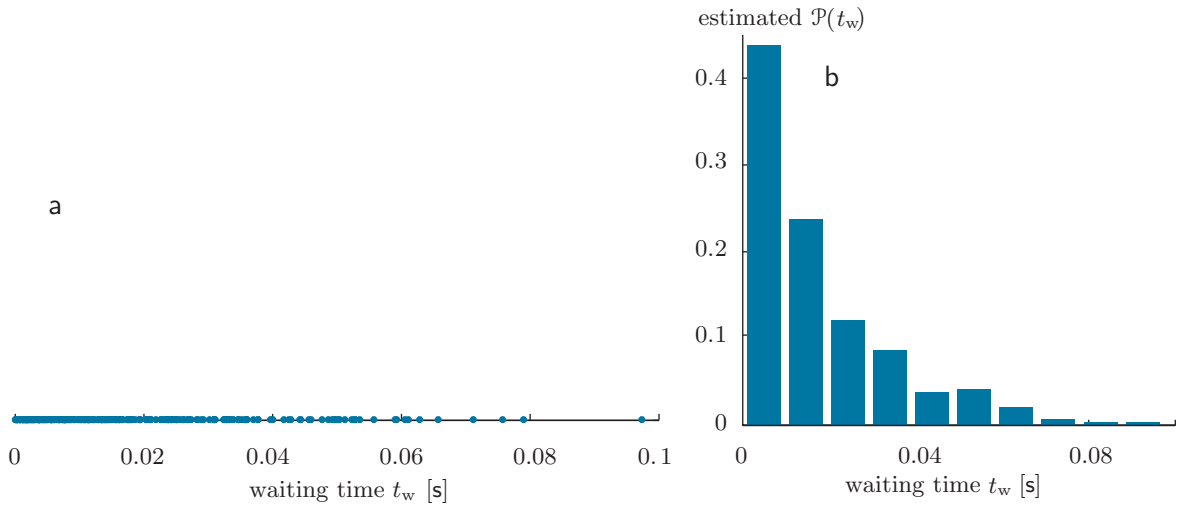


Figure 3.2 [Experimental data.] **Two ways to visualize a distribution.** (a) Cloud diagram showing the waiting times between 290 successive light detector blips as a cloud of 289 dots. The dot density is higher at the left of the diagram. (b) The same data, presented as a histogram. Taller bars correspond to greater density of dots in (a). The data have been subdivided into 10 discrete bins. [Data courtesy John F Beausang (see Dataset 3).]

times and finding their average, which is related to the intensity of the light source. But there is a limit to the information we can obtain in this way. Once we know that the general form of the distribution is Exponential, then it is completely characterized by the average waiting time; there is nothing more that any number of initial measurements can tell us about the next one, apart from refining our estimate of that one quantity.

5a. The positions of a particle undergoing Brownian motion reflect the unimaginably complex impacts of many water molecules during the intervals between snapshots ("video frames"), separated by equal time intervals Δt. Nevertheless, again we know at least a little bit about them: Such a particle will never jump, say, 1 cm from one video frame to the next, although there is no limit to how far it could move if given enough time. That is, position at video frame i does give us some partial information about position at frame $i + 1$: Successive observed positions are "correlated," in contrast to examples **1–3**.[5] However, once we know the position at frame i, then also knowing it at any frame *prior* to i gives us no additional help predicting it at $i + 1$; we say the Brownian particle "forgets" everything about its past history other than its current position, then takes a random step whose distribution of outcomes depends only on that one datum. A random system that generates a series of steps with this "forgetful" property is called a **Markov process**.

Brownian motion is a particularly simple kind of Markov process, because the distribution of positions at frame $i + 1$ has a simple form: It's a universal function, common to all steps, simply *shifted* by the position at i. Thus, if we subtract the vector position $\boldsymbol{x}_{i-1}$ from $\boldsymbol{x}_i$, the resulting displacements $\Delta\boldsymbol{x}_i$ are *uncorrelated*, with a distribution peaked at zero displacement (see Figure 3.3).

[5] Section 3.4.1 will give a more precise definition of correlation.

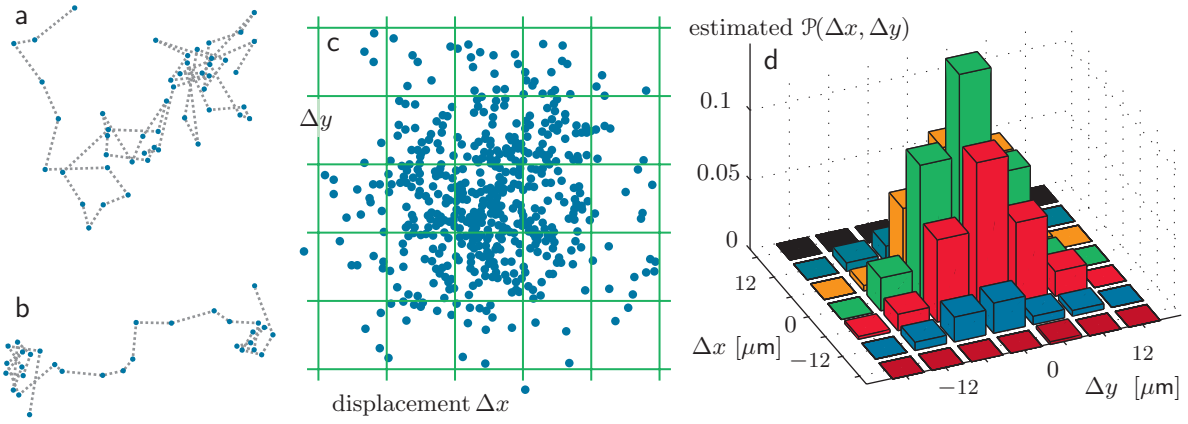


Figure 3.3 [Experimental data.] **Progressively more abstract representations of Brownian motion.** (a) *Dots* show successive observations of the position of a micrometer-size particle, taken at 30-second intervals. The *lines* merely join the dots; the particle did not really undergo straight-line motion between observations. (b) Similar data from another trial. (c) Cloud representation of the displacement vectors $\Delta \boldsymbol{x}$ joining successive positions, for 508 such position observations. Thus, a dot at the exact center would correspond to zero displacement. The grid lines are separated by about 6 μm. Thus, on a few occasions, the particle was observed to have moved more than 16 μm, though it usually moved much less than that. (d) The same data, presented as a histogram. The data have been subdivided into 49 discrete bins. [Data from Perrin, 1909; see Dataset 4.]

We will return to these key examples often as we study more biologically relevant systems. For now, we simply note that they motivate a pragmatic definition of randomness:

> *A system that yields a series of outcomes is effectively random if a list of those outcomes has no discernible, relevant structure beyond what we explicitly state.* (3.2)

The examples above described quantities that are, for many purposes, effectively random after acknowledging these characteristics:

1b–3b. Coin flips (or computer-generated random integers) are characterized by a finite list of allowed values, and the Uniform distribution on that list.

4b. Blip waiting times are characterized by a range of allowed values ($t_w > 0$), and a particular distribution on that range (in this case, not Uniform).

5b. Brownian motion is characterized by a non-Uniform distribution of positions in each video frame, which depends in a specific way on the position in the previous frame.

In short, each random system that we study has its own structure, which characterizes "what kind of randomness" it has.

The definition in Idea 3.2 may sound suspiciously imprecise. But the alternative—a precise-sounding mathematical definition—is often not helpful. How do we know for sure that any particular biological or physical system really fits the precise definition? Maybe there is some unsuspected seasonal fluctuation, or some slow drift in the electricity powering the apparatus. In fact, very often in science, a tentative identification of a supposedly random system turns out to omit some structure hiding in the actual observations (for example, correlations). Later we may discern that extra structure, and discover something new. It's best to avoid the illusion that we know everything about a system, and treat all our statements

about the kind of randomness in a system as provisional, to be sharpened as more data become available. Later sections will explain how to do this in practice.

3.2.2 Computer simulation of a random system

In addition to the integer random function mentioned earlier, any mathematical software system has another function that simulates a sample from the continuous Uniform distribution in the range between 0 and 1. We can use it to simulate a Bernoulli trial (coin flip) by drawing such a random number and comparing it to a constant ξ; if it's less than ξ, we can call the outcome heads, and otherwise tails. That is, we can partition the interval from 0 to 1 into two subintervals and report which one contains the number drawn. The probability of each outcome is the width of the corresponding subinterval.

Later chapters will extend this idea considerably, but already you can gain some valuable insight by writing simple simulations and graphing the results.[6]

3.2.3 Biological and biochemical examples

Here are three more examples with features similar to the ones in Section 3.2.1:

1c. Many organisms are **diploid**; that is, each of their cells contains two complete copies of the genome. One copy comes from the male parent, the other from the female parent. Each parent forms germ cells (egg or sperm/pollen) via **meiosis**, in which *one* copy of each gene ends up in each germ cell. That copy is essentially chosen at random from the two that were initially present. That is, for many purposes, inheritance can be thought of as a Bernoulli trial with $\xi = 1/2$, where heads could mean that the germ cell receives the copy originally given by the grandmother, and tails means the other copy.[7]

4c. Many chemical reactions can be approximated as "well mixed." In this case, the probability that the reaction will take a step in a short time interval dt depends only on the total number of reactant molecules present at the start of that interval, and not on any earlier history of the system.[8] For example, a single **enzyme** molecule wanders in a bath of other molecules, most of them irrelevant to its function. But when a particular molecular species, the enzyme's **substrate**, comes near, the enzyme can bind it, transform it chemically, and release the resulting **product** without any net change to itself. Over a time interval in which the enzyme does not significantly change the ambient concentrations of substrate, the individual appearances of product molecules have a distribution of waiting times similar to that in Figure 3.2b.

5c. Over longer times, or if the initial number of substrate molecules is not huge, it may be necessary to account for changes in population. Nevertheless, in well-mixed solutions each reaction step depends only on the *current* numbers of substrate and product molecules, not on the prior history. Even huge systems of many simultaneous reactions, among many species of molecules, can often be treated in this way. Thus, many biochemical reaction networks have the same Markov property as in example **5a** on page 38.

---

[6] See Problems 3.3 and 3.5.

[7] T_2 This simplified picture gets complicated by other genetic processes, including gene transposition, duplication, excision, and point mutation. In addition, the Bernoulli trials corresponding to different genes are not necessarily independent of one another, due to physical linkage of the genes on chromosomes.

[8] Chapter 1 assumed that this was also the case for clearance of HIV virions by the immune system.

3.2.4 False patterns: Clusters in epidemiology

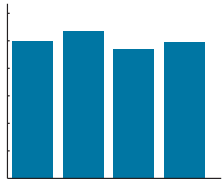


Figure 3.1a (page 37)

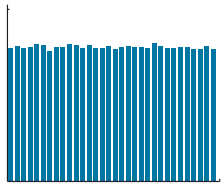


Figure 3.1b (page 37)

Look again at Figure 3.1. These figures show estimated probabilities from finite samples of data known to have been drawn from Uniform distributions. If we were told that these figures represented experimental data, however, we might wonder whether there is additional structure present. Is the extra height in the second bar in panel (a) significant?

Human intuition is not always a reliable guide to questions like this one when the available data are limited. For example, we may have the home addresses of a number of people diagnosed with a particular disease, and we may observe an apparent geographical cluster in those addresses. But even Uniformly distributed points will show apparent clusters, if our sample is not very large. Chapter 4 will develop some tools to assess questions like these.[9]

3.3 Probability Distribution of a Discrete Random System

3.3.1 A probability distribution describes to what extent a random system is, and is not, predictable

Our provisional definition of randomness, Idea 3.2, hinged on the idea of "structure." To make this idea more precise, note that the examples given so far all yield measurements that are, at least in principle, **replicable**. That is, they come from physical systems that are simple enough to reproduce in many copies, each one identical to the others in all relevant aspects and each unconnected to all the others. Let's recall some of the examples in Section 3.2.1:

1d. We can make many identical coins and flip them all using similar hand gestures.

4d. We can construct many identical light sources and shine them on identical light detectors.

5d. We can construct many chambers, each of which releases a micrometer-size particle at a particular point at time zero and observes its subsequent Brownian motion.

We can then use such repeated measurements to learn what discernible, relevant structure our random system may have. Here, "structure" means *everything that we can learn from a large set of measurements that can help us to predict the next one.* Each example given had a rather small amount of structure in this sense:

1e. All that we can learn from a large number of coin tosses that's helpful for guessing the result of the next one is the single number ξ characterizing a Bernoulli trial.

2e–3e. Similarly in these examples, once we list the allowed outcomes and determine that each is equally probable (Uniform distribution), nothing more can be gleaned from past outcomes.

4e. For light detection, each blip waiting time is independent of the others, so again, the distribution of those times is all that we can measure that is useful for predicting the next one. Unlike examples **2e** and **3e**, however, in this case the distribution is non-Uniform (Figure 3.2).

5e. In Brownian motion, the successive positions of a particle are not independent. However, we may still take the *entire trajectory* of the particle throughout our trial as the "outcome" (or observation), and assert that each is independent of the other trials. Then, what we can say about the various trajectories is that the ones with lots

[9] See Problem 4.10.

of large jumps are less probable than the ones with few: There is a distribution on the *space of trajectories*. It's not easy to visualize such a high-dimensional distribution, but Figure 3.3 simplified it by looking only at the net displacement $\Delta \boldsymbol{x}$ after a particular elapsed time.

With these intuitive ideas in place, we can now give a formal definition of a random system's "structure," or probability distribution, starting with the special case of discrete outcomes.

Suppose that a random system is replicable; that is, it can be measured repeatedly, independently, and under the same conditions. Suppose also that the outcomes are always items drawn from a discrete list, indexed by ℓ.[10] Suppose that we make many measurements of the outcome (N_{tot} of them) and find that $\ell = 1$ on N_1 occasions, and so on. Thus, N_ℓ is an integer, the number of times that outcome ℓ was observed; it's called the **frequency** of outcome ℓ.[11] If we start all over with another N_{tot} measurements, we'll get different frequencies N'_ℓ, but for large enough N_{tot} they should be about equal to the corresponding N_ℓ. We say that the discrete **probability distribution** of the outcome ℓ is the fraction of trials that yielded ℓ, or[12]

$$\mathcal{P}(\ell) = \lim_{N_{\text{tot}} \to \infty} N_\ell / N_{\text{tot}}. \tag{3.3}$$

Note that $\mathcal{P}(\ell)$ is always nonnegative. Furthermore,

Any discrete probability distribution function is dimensionless,

because for any ℓ, $\mathcal{P}(\ell)$ involves the ratio of two *integers*. The N_ℓ's must add up to N_{tot} (every observation is assigned to *some* outcome). Hence, any discrete distribution must have the property that

$$\sum_\ell \mathcal{P}(\ell) = 1. \quad \text{normalization condition, discrete case} \tag{3.4}$$

Equation 3.3 can also be used with a finite number of observations, to obtain an *estimate* of $\mathcal{P}(\ell)$. When drawing graphs, we often indicate such estimates by representing the values by bars. Then the heights of the bars must all add up to 1, as they do in Figures 3.1a,b, 3.2b, and 3.3d. This representation looks like a histogram, and indeed it differs from an ordinary histogram only in that each bar has been scaled by $1/N_{\text{tot}}$.

We'll call the list of all possible distinct outcomes the **sample space**. If ℓ_1 and ℓ_2 are two distinct outcomes, then we may also ask for the probability that "either ℓ_1 or ℓ_2 was observed." More generally, an **event** E is any subset of the sample space, and it "occurs" whenever we draw from our random system an outcome ℓ that belongs to the subset E. The

---

[10] That list may be infinite, but "discrete" means that we can at least label the entries by an integer. Chapter 5 will discuss continuous distributions.

[11] There is an unavoidable collision of language associated with this term from probability. ΔN is an integer, with no units. But in physics, "frequency" usually refers to a different sort of quantity, with units $\mathbb{T}^{-1}$ (for example, the frequency of a musical note). We must rely on context to determine which meaning is intended.

[12] $\boxed{T_2}$ Some authors call $\mathcal{P}(\ell)$ a **probability mass function** and, unfortunately, assign a completely different meaning to the words "probability distribution function." (That meaning is what we will instead call the "cumulative distribution.")

probability of an event E is the fraction of all draws that yield an outcome belonging to E. The quantity $\mathcal{P}(\ell)$ is just the special case corresponding to an event containing only one point in sample space.

We can also regard events as *statements:* The event E corresponds to the statement "The outcome was observed to be in the set E," which will be true or false every time we draw from (observe) the random system. We can then interpret the logical operations **or**, **and**, and **not** as the usual set operations of union, intersection, and complement.

3.3.2 A random variable has a sample space with numerical meaning

Section 3.3.1 introduced a lot of abstract jargon; let's pause here to give some more concrete examples.

A Bernoulli trial has a sample space with just two points, so $\mathcal{P}_{\text{bern}}(\ell)$ consists of just two numbers, namely, $\mathcal{P}_{\text{bern}}(\underline{\text{heads}})$ and $\mathcal{P}_{\text{bern}}(\underline{\text{tails}})$. We have already set the convention that

$$\mathcal{P}_{\text{bern}}(\underline{\text{heads}};\xi) = \xi \quad \text{and} \quad \mathcal{P}_{\text{bern}}(\underline{\text{tails}};\xi) = (1-\xi), \tag{3.5}$$

where ξ is a number between 0 and 1. The semicolon followed by ξ reminds us that "the" Bernoulli trial is really a *family* of distributions depending on the *parameter* ξ. Everything before the semicolon specifies an outcome; everything after is a parameter. For example, the fair-coin flip is the special case $\mathcal{P}_{\text{bern}}(s;1/2)$, where the outcome label s is a variable that ranges over the two values {$\underline{\text{heads}}$, $\underline{\text{tails}}$}.

We may have an interpretation for the sample space in which the outcome label is literally a number (like the number of petals on a daisy flower or the number of copies of an RNA molecule in a cell at some moment). Or it may simply be an index to a list of outcomes without any special numerical significance; for example, our random system may consist of successive cases of a disease, and the outcome label s indexes a list of towns where the cases were found. Even in such a situation, there may be one or more interesting numerical *functions of* s. For example, $f(s)$ could be the distance of each town from a particular point, such as the location of a power-generating station. Any such numerical function on the sample space is called a **random variable**. If the outcome label itself can be naturally interpreted as a number, we'll usually call it ℓ; in this case, $f(\ell)$ is any ordinary function, or even ℓ itself.

We already know a rather dull example of a random variable: the Uniformly distributed discrete random variable on some range (Figure 3.1a,b). For example, if ℓ is restricted to integer values between 3 and 6, then this Uniform distribution may be called $\mathcal{P}_{\text{unif}}(\ell;3,6)$. It equals $1/4$ if $\ell = 3, 4, 5,$ or 6, and 0 otherwise. The semicolon again separates the potential value of a random variable (here, ℓ) from some parameters specifying which distribution function in the family is meant.[13] We will eventually define several such parametrized families of idealized distribution functions.

Another example, which we'll encounter in later chapters, is the "Geometric distribution." To motivate it, imagine a frog that strikes at flies, not always successfully. Each attempt is an independent Bernoulli trial with some probability of success ξ. How long must the frog wait for its next meal? Clearly there is no unique answer to this question, but we can nevertheless ask about the *distribution* of answers. Letting j denote the number of attempts

[13] Often we will omit parameter values to shorten our notation, for example, by writing $\mathcal{P}_{\text{unif}}(\ell)$ instead of the cumbersome $\mathcal{P}_{\text{unif}}(\ell;3,6)$, if the meaning is clear from context.

needed to get the next success, we'll call this distribution $\mathcal{P}_{\text{geom}}(j;\xi)$. Section 3.4.1.2 will work out an explicit formula for this distribution. For now just note that, in this case, the random variable j can take any positive integer value—the sample space is discrete (although infinite). Also note that, as with the previous examples, "the" Geometric distribution is really a *family* depending on the value of a parameter ξ, which can be any number between 0 and 1.

3.3.3 The addition rule

The probability that the next measured value of ℓ is *either* ℓ_1 *or* ℓ_2 is simply $\mathcal{P}(\ell_1) + \mathcal{P}(\ell_2)$ (unless $\ell_1 = \ell_2$). More generally, if two events E_1 and E_2 have no overlap, we say that they are **mutually exclusive**; then Equation 3.3 implies that

$$\mathcal{P}(\mathsf{E}_1 \textbf{ or } \mathsf{E}_2) = \mathcal{P}(\mathsf{E}_1) + \mathcal{P}(\mathsf{E}_2). \qquad \textbf{addition rule} \text{ for mutually exclusive events} \tag{3.6}$$

If the events do overlap, then just adding the probabilities will overstate the probability of (E_1 **or** E_2), because some outcomes will be counted twice.[14] In this case, we must modify our rule to say that

$$\mathcal{P}(\mathsf{E}_1 \textbf{ or } \mathsf{E}_2) = \mathcal{P}(\mathsf{E}_1) + \mathcal{P}(\mathsf{E}_2) - \mathcal{P}(\mathsf{E}_1 \textbf{ and } \mathsf{E}_2). \quad \text{general } \textbf{addition rule} \tag{3.7}$$

Your Turn 3A

Prove Equation 3.7 starting from Equation 3.3.

3.3.4 The negation rule

Let **not**-E be the statement that "the outcome is not included in event E." Then E and **not**-E are mutually exclusive, and, moreover, either one or the other is true for every outcome. In this case, Equation 3.3 implies that

$$\mathcal{P}(\textbf{not-}\mathsf{E}) = 1 - \mathcal{P}(\mathsf{E}). \quad \textbf{negation rule} \tag{3.8}$$

This obvious-seeming rule can be surprisingly helpful when we want to understand a complex event.[15]

If E is one of the outcomes in a Bernoulli trial, then **not**-E is the other one, and Equation 3.8 is the same as the normalization condition. More generally, suppose that we have many events $\mathsf{E}_1, \ldots, \mathsf{E}_n$ with the property that any two are mutually exclusive. Also suppose that together they cover the entire sample space. Then Equation 3.8 generalizes to the statement that the sum of all the $\mathcal{P}(\mathsf{E}_i)$ equals one—a more general form of the normalization condition. For example, each of the bars in Figure 3.2b corresponds to an

[14] In logic, "E_1 **or** E_2" means either E_1, E_2, *or both*, is true.
[15] See Problem 3.13.

event defined by a range of possible waiting times; we say that we have **binned the data**, converting a lot of observations of a continuous quantity into a discrete set of bars, whose heights must sum to 1.[16]

3.4 Conditional Probability

3.4.1 Independent events and the product rule

Consider two scenarios:

a. A friend rolls a six-sided die but doesn't show you the result, and then asks you if you'd like to place a bet that wins if the die landed with 5 facing up. Before you reply, a bystander comes by and adds the information that the die is showing some *odd* number. Does this change your assessment of the risk of the bet?

b. A friend rolls a die and flips a coin, doesn't show you the results, and then asks you if you'd like to place a bet that wins if the die landed with 5 facing up. Before you reply, the friend suddenly adds the information that the coin landed with heads up. Does this change your assessment of the risk of the bet?

The reason you changed your opinion in scenario **a** is that the additional information you gained eliminated some of the sample space (all the outcomes corresponding to even numbers). If we roll a die many times but disregard all rolls that came up even, then Equation 3.3 says that the probability of rolling 5, *given that we rolled an odd number*, is 1/3. Letting E_5 be the event "roll a 5" and E_{odd} the event "roll an odd number," we write this quantity as $\mathcal{P}(\mathsf{E}_5 \mid \mathsf{E}_{\text{odd}})$ and call it "the conditional probability of E_5 given E_{odd}." More generally, the conditional probability $\mathcal{P}(\mathsf{E} \mid \mathsf{E}')$ accounts for partial information by restricting the denominator in Equation 3.3 to only those measurements for which E' is true and restricting the numerator to only those measurements for which *both* E and E' are true:

$$\mathcal{P}(\mathsf{E} \mid \mathsf{E}') = \lim_{N_{\text{tot}} \to \infty} \frac{N(\mathsf{E} \textbf{ and } \mathsf{E}')}{N(\mathsf{E}')}.$$

We can give a useful rule for computing conditional probabilities by dividing both numerator and denominator by the same thing, the total number of all measurements made:

$$\mathcal{P}(\mathsf{E} \mid \mathsf{E}') = \lim_{N_{\text{tot}} \to \infty} \frac{N(\mathsf{E} \textbf{ and } \mathsf{E}')/N_{\text{tot}}}{N(\mathsf{E}')/N_{\text{tot}}}, \text{ or} \tag{3.9}$$

$$\mathcal{P}(\mathsf{E} \mid \mathsf{E}') = \frac{\mathcal{P}(\mathsf{E} \textbf{ and } \mathsf{E}')}{\mathcal{P}(\mathsf{E}')}. \quad \textbf{conditional probability} \tag{3.10}$$

Equivalently,

$$\mathcal{P}(\mathsf{E} \textbf{ and } \mathsf{E}') = \mathcal{P}(\mathsf{E} \mid \mathsf{E}') \times \mathcal{P}(\mathsf{E}'). \quad \text{general } \textbf{product rule} \tag{3.11}$$

[16] T_2 Binning isn't always necessary nor desirable; see Section 6.2.4′ (page 142).

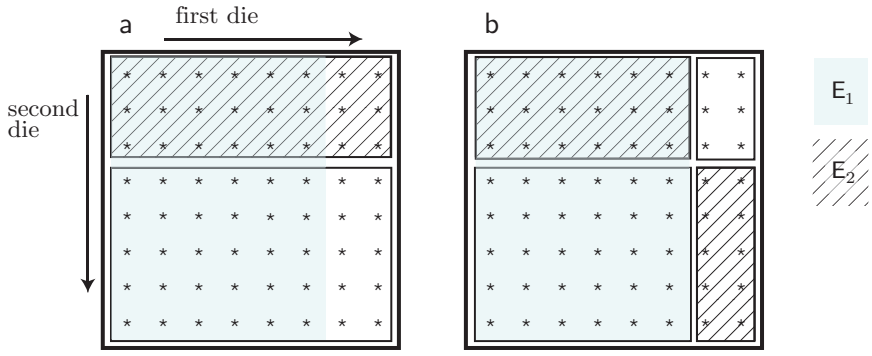


Figure 3.4 [Box diagrams.] **Graphical representations of joint probability distributions.** Each panel represents a distribution graphically as a partitioning of a square. Each consists of 64 equally probable outcomes. (a) Within this sample space, event E_1 corresponds to rolls of two eight-sided dice for which the number on the first die was less than 7; those events lie in the *colored part* of the square. Event E_2 corresponds to rolls for which the second number was less than 4; those events are represented by the *hatched region*. In this situation, E_1 and E_2 are statistically independent. We can see this geometrically because, for example, (E_1 **and** E_2) occupies the upper left rectangle, whose width is that of E_1 and height is that of E_2. (b) A different choice of two events in the same sample space. (E_1 **and** E_2) again occupies the upper left rectangle, but this time its area is *not* the product of $\mathcal{P}(\mathsf{E}_1)$ and $\mathcal{P}(\mathsf{E}_2)$. Thus, these two events are not independent.

A special case of this rule is particularly important. Sometimes knowing that E' is true tells us nothing relevant to predicting E. (That's why you didn't change your bet in scenario **b** above.) That is, suppose that the additional information you were given was *irrelevant* to what you were trying to predict: $\mathcal{P}(\mathsf{E}_5 \mid \mathsf{E}_{\underline{\text{heads}}}) = \mathcal{P}(\mathsf{E}_5)$. The product rule then implies that, in this case, $\mathcal{P}(\mathsf{E}_5 \textbf{ and } \mathsf{E}_{\underline{\text{heads}}}) = \mathcal{P}(\mathsf{E}_5) \times \mathcal{P}(\mathsf{E}_{\underline{\text{heads}}})$. More generally, we say that two events are **statistically independent**[17] if

$$\mathcal{P}(\mathsf{E} \textbf{ and } \mathsf{E}') = \mathcal{P}(\mathsf{E}) \times \mathcal{P}(\mathsf{E}'). \quad \text{statistically independent events} \tag{3.12}$$

Two events that are *not* statistically independent are said to be **correlated**.

Equation 3.12 is very useful because often we have a physical model of a random system that states a priori that two events are independent.

It's good to have a pictorial representation of any abstract concept. To represent a random system, we can draw a unit square (sides of length 1), then divide it into boxes corresponding to all of the possible outcomes. For example, suppose that we roll a pair of fair eight-sided dice, so that our sample space consists of 64 elementary events, each of which is equally probable. Figure 3.4a shows the elementary events as asterisks. Symbols set in a shaded background correspond to an event we'll call E_1; the unshaded region is **not**-E_1. The hatched region corresponds to another event called E_2; its complement is **not**-E_2. Because every outcome has the same probability, the probabilities of various events are simply the number of outcomes they contain, times 1/64; equivalently, the probabilities correspond to the *areas* of the various regions in the unit square.

In Figure 3.4a, both blocks on the left side are colored; both on the right are not. Both blocks on the top are hatched; both on the bottom are not. This arrangement implies

[17] The abbreviation "independent," or the synonym "uncorrelated," is frequently used instead of "statistically independent."

that $\mathcal{P}(\mathsf{E}_1 \textbf{ and } \mathsf{E}_2) = \mathcal{P}(\mathsf{E}_1)\mathcal{P}(\mathsf{E}_2)$, and similarly for the other three blocks. Thus, the joint distribution has the product form that implies independence according to the product rule. In contrast, Figure 3.4b graphically represents a different arrangement, in which the events are not independent.

Your Turn 3B

Make this argument more explicit. That is, calculate $\mathcal{P}(\mathsf{E}_1)$ and $\mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2)$ for each of Figures 3.4a,b, and comment.

T_2 *Section 3.4.1′ (page 60) develops more general forms of the product and negation rules.*

3.4.1.1 Crib death and the prosecutor's fallacy

Officials in the United Kingdom prosecuted hundreds of women, mainly in the 1990s, for the murder of their own infants, who died in their cribs. The families of the targeted women had suffered multiple crib deaths, and the arguments made to juries often took the form that "one is a tragedy, two is suspicious, and three is murder." In one case, an expert justified this claim by noting that, at that time, about one infant in 8500 died in its crib for no known cause in the United Kingdom. The expert then calculated the probability of two such deaths occurring naturally in a family as $(1/8500)^2$, which is a tiny number.

It is true that the observed occurrence of multiple crib deaths in one family, in a population of fewer than 8500^2 families, strongly suggests that successive instances are not statistically independent. The logical flaw in the argument is sometimes called the "prosecutor's fallacy"; it lies in the assumption that *the only possible source* of this nonindependence is willful murder. For example, there could instead be a genetic predisposition to crib death, a noncriminal cause that would nevertheless be correlated within families. After an intervention from the Royal Statistical Society, the UK attorney general initiated legal review of every one of the 258 convictions.

Crib death could also be related to ignorance or custom, which tends to remain constant within each family (hence correlated between successive children). Interestingly, after a vigorous informational campaign to convince parents to put their babies to sleep on their back or side, the incidence of crib death in the United Kingdom dropped by 70%. Had the earlier crib deaths been cases of willful murder, it would have been a remarkable coincidence that they suddenly declined at exactly the same time as the information campaign!

3.4.1.2 The Geometric distribution describes the waiting times for success in a series of independent trials

Section 3.3.2 introduced a problem (frog striking at flies) involving repeated, independent attempts at a yes/no goal, each with probability ξ of success. Let j denote the number of attempts made from one success to the next. For example, $j = 2$ means one failure followed by success (two attempts in all). Let's find the probability distribution of the random variable j.

Once a fly has been caught, there is probability ξ of succeeding again, on the very next attempt: $\mathcal{P}_{\text{geom}}(1;\xi) = \xi$. The outcome of exactly one failure followed by success is then a product: $\mathcal{P}_{\text{geom}}(2;\xi) = \xi(1-\xi)$, and so on. That is,

$$\mathcal{P}_{\text{geom}}(j;\xi) = \xi(1-\xi)^{j-1}, \text{ for } j = 1, 2, \ldots. \quad \textbf{Geometric distribution} \tag{3.13}$$

This family of discrete probability distribution functions is called "Geometric" because each value is a constant times the previous one—a geometric sequence.

Your Turn 3C

a. Graph this probability distribution function for fixed values of $\xi = 0.15,\ 0.5$, and 0.9. Because the function $\mathcal{P}_{\text{geom}}(j)$ is defined only at integer values of j, be sure to indicate this by drawing dots or some other symbols at each point—not just a set of line segments.
b. Explain the features of your graphs in terms of the underlying situation being described. What is the most probable value of j in each graph? Think about why you got that result.

Your Turn 3D

Because the values $j = 1,\ 2,\ \ldots$ represent a complete set of mutually exclusive possibilities, we must have that $\sum_{j=1}^{\infty} \mathcal{P}_{\text{geom}}(j;\xi) = 1$ for any value of ξ. Confirm this by using the Taylor series for the function $1/(1-x)$, evaluated near $x = 0$ (see page 19).

You'll work out the basic properties of this family of distributions in Problem 7.2.

3.4.2 Joint distributions

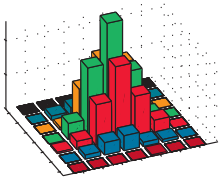


Figure 3.3d (page 39)

Sometimes a random system yields measurements that each consist of *two* pieces of information; that is, the system's sample space can be naturally labeled by pairs of discrete variables. Consider the combined act of rolling an ordinary six-sided die and also flipping a coin. The sample space consists of all pairs (ℓ, s), where ℓ runs over the list of all allowed outcomes for the die and s runs over those for the coin. Thus, the sample space consists of a total of 12 outcomes. The probability distribution $\mathcal{P}(\ell, s)$, still defined by Equation 3.3, is called the **joint distribution** of ℓ and s. It can be thought of as a table, whose entry in row ℓ and column s is $\mathcal{P}(\ell, s)$. Two-dimensional Brownian motion is a more biological example: Figure 3.3d shows the joint distribution of the random variables Δx and Δy, the components of the displacement vector $\Delta \boldsymbol{x}$ after a fixed elapsed time.

Your Turn 3E

Suppose that we roll two six-sided dice. What's the probability that the numbers on the dice add up to 2? To 6? To 12? Think about how you used both the addition and product rules for this calculation.

We may not be interested in the value of s, however. In that case, we can usefully reduce the joint distribution by considering the event $\mathsf{E}_{\ell=\ell_0}$, which is the statement that the random system generated any outcome for which ℓ has the particular value ℓ_0, with no restriction on s. The probability $\mathcal{P}(\mathsf{E}_{\ell=\ell_0})$ is often written simply as $\mathcal{P}_\ell(\ell_0)$, and is called the **marginal distribution** over s; we also say that we obtained it from the joint distribution by "marginalizing" s.[18] We implicitly did this when we reduced the entire path of Brownian motion over 30 s to just the final displacement in Figure 3.3.

[18] The subscript "ℓ" serves to distinguish this distribution from other functions of one variable, for example, the distribution $\mathcal{P}_s$ obtained by marginalizing ℓ. When the meaning is clear, we may sometimes drop this subscript. The notation $\mathcal{P}(\ell_0, s_0)$ does not need any subscript, because (ℓ_0, s_0) completely specify a point in the coin/die sample space.

We may find that each of the events $\mathsf{E}_{\ell=\ell_0}$ is statistically independent of each of the $\mathsf{E}_{s=s_0}$. In that case, we say that the random variables ℓ and s are themselves independent.

Your Turn 3F

a. Show that $\mathcal{P}_\ell(\ell_0) = \sum_s \mathcal{P}(\ell_0, s)$.
b. If ℓ and s are independent, then show that $\mathcal{P}(\ell, s) = \mathcal{P}_\ell(\ell) \times \mathcal{P}_s(s)$.
c. Imagine a random system in which each "observation" involves drawing a card from a shuffled deck, and then, without replacing it, drawing a second card. If ℓ is the first card's name and s the second one's, are these independent random variables?

The next idea is simple, but subtle enough to be worth stating carefully. Suppose that ℓ corresponds to the roll of a four-sided die and s to a coin flip. We often want to sum over all the possibilities for ℓ, s—for example, to check normalization or compute some average. Let's symbolically call the terms of the sum $[\ell, s]$. We can group the sum in two ways:

$$\Big([1, \underline{\text{tails}}]+[2, \underline{\text{tails}}]+[3, \underline{\text{tails}}]+[4, \underline{\text{tails}}]\Big)+\Big([1, \underline{\text{heads}}]+[2, \underline{\text{heads}}]+[3, \underline{\text{heads}}]+[4, \underline{\text{heads}}]\Big)$$

or

$$\Big([1, \underline{\text{tails}}] + [1, \underline{\text{heads}}]\Big) + \Big([2, \underline{\text{tails}}] + [2, \underline{\text{heads}}]\Big) + \Big([3, \underline{\text{tails}}] + [3, \underline{\text{heads}}]\Big) + \Big([4, \underline{\text{tails}}] + [4, \underline{\text{heads}}]\Big).$$

Either way, it's the same eight terms, just grouped differently. But one of these versions may make it easier to see a point than the other, so often it's helpful to try both.

The first formula above can be expressed in words as "Hold s fixed to $\underline{\text{tails}}$ while summing ℓ, then hold s fixed to $\underline{\text{heads}}$ while again summing ℓ." The second formula can be expressed as "Hold ℓ fixed to 1 while summing s, and so on." The fact that these recipes give the same answer can be written symbolically as

$$\sum_{\ell,s}\Big(\cdots\Big) = \sum_s\Big(\sum_\ell(\cdots)\Big) = \sum_\ell\Big(\sum_s(\cdots)\Big). \tag{3.14}$$

Use this insight to work the following problem:

Your Turn 3G

a. Show that the joint distribution for two independent sets of outcomes will automatically be correctly normalized if the two marginal distributions (for example, our $\mathcal{P}_{\text{die}}$ and $\mathcal{P}_{\text{coin}}$) each have that property.
b. This time, suppose that we are given a properly normalized joint distribution, *not* necessarily for independent outcomes, and we compute the marginal distribution by using the formula you found in Your Turn 3F. Show that the resulting $\mathcal{P}_\ell(\ell)$ is automatically properly normalized.

3.4.3 The proper interpretation of medical tests requires an understanding of conditional probability

Statement of the problem

Let's apply these ideas to a problem whose solution surprises many people. This problem *can* be solved accurately by using common sense, but many people perceive alternate, wrong solutions to be equally reasonable. The concept of conditional probability offers a more sure-footed approach to problems of this sort.

Suppose that you have been tested for some dangerous disease. You participated in a mass random screening; you do not feel sick. The test comes back "positive," that is, indicating that you in fact have the disease. Worse, your doctor tells you the test is "97% accurate." That sounds bad.

Situations like this one are very common in science. We measure something; it's not precisely what we wanted to know, but neither is it irrelevant. Now we must attempt an *inference:* What can we say about the question of interest, based on the available new information? Returning to the specific question, you want to know, "Am I sick?" The ideas of conditional probability let us phrase this question precisely: We wish to know $\mathcal{P}(\text{sick} \mid \text{positive})$, the *probability* of being sick, given one positive test result.

To answer the question, we need some more precise information. The accuracy of a yes/no medical test actually has two distinct components:

- The **sensitivity** is the fraction of truly sick people who test positive. A sensitive test catches almost every sick person; that is, it yields very few **false-negative** results. For illustration, let's assume that the test has 97% sensitivity (a false-negative rate of 3%).
- The **selectivity** is the fraction of truly healthy people who test negative. High selectivity means that the test gives very few **false-positive** results. Let's assume that the test also has 97% selectivity (a false-positive rate of 3%).

In practice, false-positive and -negative results can arise from human error (a label falls off a test tube), intrinsic fluctuations, sample contamination, and so on. Sometimes sensitivity and selectivity depend on a threshold chosen when setting a lab protocol, so that one of them can be increased, but only at the expense of lowering the other one.

Analysis

Let E_{sick} be the event that a randomly chosen member of the population is sick, and E_{pos} the event that a randomly chosen member of the population tests positive. Now, certainly, these two events are *not* independent—the test does tell us *something*—in fact, quite a lot, according to the data given. But the two events are not quite synonymous, because neither the sensitivity nor the selectivity is perfect. Let's abbreviate $\mathcal{P}(S) = \mathcal{P}(\mathsf{E}_{\text{sick}})$, and so on. In this language, the sensitivity is $\mathcal{P}(P \mid S) = 0.97$.

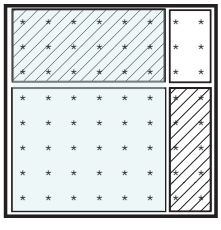


Figure 3.4b (page 46)

Before writing any more abstract formulas, let's attempt a pictorial representation of the problem. We represent the complete population by a 1×1 square containing evenly spaced points, with a point for every individual in the very large population under study. Then the probability of being in any subset is simply the *area* it fills on the square. We segregate the population into four categories based on sick/healthy status (S/H) and test result (P/N), and give names to their areas. For example, let $\mathcal{P}_{HN}$ denote $\mathcal{P}(\text{healthy } \textbf{and} \text{ negative result})$, and so on. Because the test result is not independent of the health of the patient, the figure is similar to Figure 3.4b.

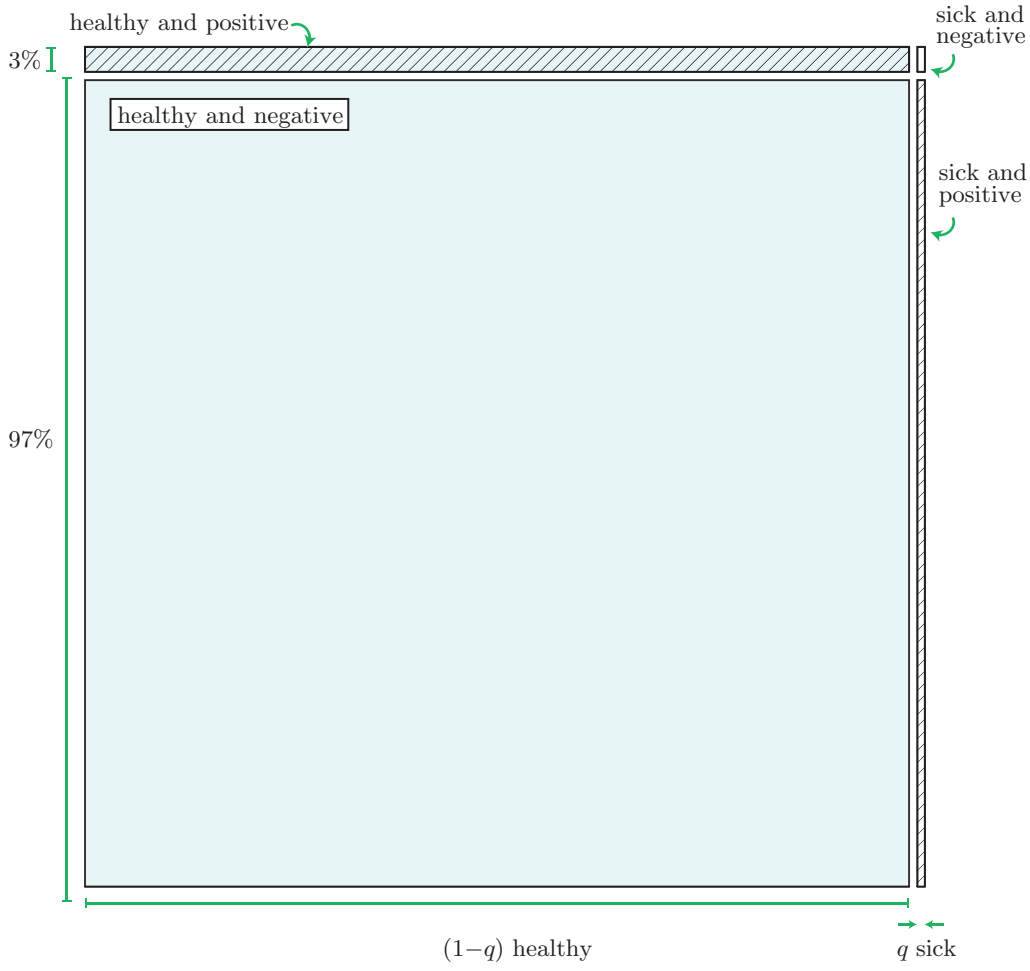


Figure 3.5 [Box diagram.] **A joint distribution representing a medical test.** The labels refer to healthy (*colored*) versus sick patients, and to positive (*hatched*) versus negative test results. The events E_{sick} and E_{pos} are highly (though not perfectly) correlated, so the labels resemble those in Figure 3.4b, not Figure 3.4a (page 46).

Figure 3.5 illustrates the information we have been given. It is partitioned horizontally in such a way that

$$\text{sensitivity} = \mathcal{P}(P \mid S) = \frac{\mathcal{P}(S \textbf{ and } P)}{\mathcal{P}(S)} = \frac{\mathcal{P}_{SP}}{\mathcal{P}_{SN} + \mathcal{P}_{SP}} = 97\%.$$

Your Turn 3H

Confirm that the figure also depicts a selectivity of 97%.

In our imagined scenario, you know you tested positive, but you wish to know whether you're sick. The probability, given what you know, is then

$$\mathcal{P}(S \mid P) = \frac{\mathcal{P}(S \textbf{ and } P)}{\mathcal{P}(P)} = \frac{\mathcal{P}_{SP}}{\mathcal{P}_{SP} + \mathcal{P}_{HP}}. \tag{3.15}$$

So, are you sick? Perhaps surprisingly, there is *no way to answer with the given information*. Figure 3.5 makes it clear that one additional, crucial bit of information is still missing: the fraction of the overall population that is sick. Suppose that you go back to your doctor and find that it's $\mathcal{P}(S) = 0.9\%$. This quantity is called q in the figure. Then $\mathcal{P}_{HP} = 0.03 \times (1 - q)$ and so on, and we can finish evaluating Equation 3.15:

$$\mathcal{P}(S \mid P) = \frac{0.97 \times 0.009}{0.97 \times 0.009 + 0.03 \times 0.991} = \left[1 + \frac{0.03 \times 0.991}{0.97 \times 0.009}\right]^{-1} \approx \frac{1}{4}. \tag{3.16}$$

Remarkably, although you tested positive, and the test was "97% accurate," you are *probably not sick*.[19]

What just happened? Suppose that we could test the *entire* population. The huge majority of healthy people would generate some false-positive results—more than the number of true positives from the tiny minority of sick people. That's the commonsense analysis. In graphical language, region *HP* of Figure 3.5 is not much smaller than the region *SP* that concerns us—instead, it's *larger* than *SP*.

3.4.4 The Bayes formula streamlines calculations involving conditional probability

Situations like the one in the previous subsection arise often, so it is worthwhile to create a general tool. Consider two events E_1 and E_2, which may or may not be independent.

Notice that (E_1 **and** E_2) is exactly the same event as (E_2 **and** E_1). Equation 3.11 (page 45) therefore implies that $\mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2) \times \mathcal{P}(\mathsf{E}_2) = \mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_1) \times \mathcal{P}(\mathsf{E}_1)$. Rearranging slightly gives

$$\mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2) = \mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_1)\frac{\mathcal{P}(\mathsf{E}_1)}{\mathcal{P}(\mathsf{E}_2)}. \quad \textbf{Bayes formula} \tag{3.17}$$

In everyday life, people often confuse the conditional probabilities $\mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2)$ and $\mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_1)$. The Bayes formula quantifies how they differ.

Equation 3.17 formalizes a procedure that we all use informally. When evaluating a claim E_1, we generally have some notion of how probable it is. We call this our **prior** assessment because it's how strongly we believed E_1 prior to obtaining some new information. After obtaining the new information that E_2 is true, we *update* our prior $\mathcal{P}(\mathsf{E}_1)$ to a new **posterior** assessment $\mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2)$. That is, the posterior is the probability of E_1, given the new information that E_2 is true. The Bayes formula tells us that we can compute the posterior in terms of $\mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_1)$, which in this context is called the **likelihood**. The formula is useful because sometimes we know the likelihood a priori.

For example, the Bayes formula lets us automate the reasoning of Section 3.4.3.[20] In this context, Equation 3.17 says that $\mathcal{P}(S \mid P) = \mathcal{P}(P \mid S)\mathcal{P}(S)/\mathcal{P}(P)$. The numerator equals the sensitivity times q. The denominator can be expressed in terms of the two ways of getting

[19] A one-in-four chance of being sick may nevertheless warrant medical intervention, or at least further testing. "Decision theory" seeks to weight different proposed courses of action according to the probabilities of various outcomes, as well as the severity of the consequences of each action.

[20] The general framework also lets us handle other situations, such as those in which selectivity is not equal to sensitivity (see Problem 3.10).

a positive test result:

$$\mathcal{P}(P) = \mathcal{P}(P \mid S)\mathcal{P}(S) + \mathcal{P}(P \mid \textbf{not–}S)\mathcal{P}(\textbf{not–}S). \tag{3.18}$$

Your Turn 3I

Prove Equation 3.18 by using Equation 3.11. Then combine it with the Bayes formula to recover Equation 3.16.

Your Turn 3J

a. Suppose that our test has *perfect* sensitivity and selectivity. Write the Bayes formula for this case, and confirm that it connects with what you expect.
b. Suppose that our test is *worthless*; that is, the events E_{sick} and E_{pos} are statistically independent. Confirm that in this case, too, the math connects with what you expect.

T2 *Section 3.4.4′ (page 60) develops an extended form of the Bayes formula.*

3.5 Expectations and Moments

Suppose that two people play a game in which each move is a Bernoulli trial. Nick pays Nora a penny each time the outcome $s =$ tails; otherwise, Nora pays Nick two pennies. A "round" consists of N_{tot} moves. Clearly, Nick can expect to win about $N_{\text{tot}}\xi$ times and lose $N_{\text{tot}}(1-\xi)$ times. Thus, he can expect his bank balance to have changed by about $N_{\text{tot}}\big(2\xi - (1-\xi)\big)$ pennies, although in every round the exact result will be different.

But players in a game of chance have other concerns besides the "typical" net outcome—for example, each will also want to know, "What is the risk of doing substantially *worse* than the typical outcome?" Other living creatures also play games like these, often with higher stakes.

3.5.1 The expectation expresses the average of a random variable over many trials

To make the questions more precise, let's begin by introducing a random variable $f(s)$ that equals $+2$ for $s =$ heads and -1 for $s =$ tails. Then, one useful descriptor of the game is the average of the values that f takes when we make N_{tot} measurements, in the limit of large N_{tot}. This quantity is called the **expectation** of f, and is denoted by the symbol $\langle f \rangle$. But we *don't* mean that we "expect" to observe this exact value in any real measurement; for example, in a discrete distribution, $\langle f \rangle$ generally falls between two allowed values of f, and so will *never* actually be observed.

Example Use Equation 3.3 to show that the expectation of f can be re-expressed by the formula

$$\langle f \rangle = \sum_s f(s)\mathcal{P}(s). \tag{3.19}$$

In this formula, the sum runs only over the list of possible outcomes (not over all N_{tot} repeated measurements); but each term is *weighted* by that outcome's probability.

Solution In the example of a coin-flipping game, suppose that N_1 of the flips yielded heads and N_2 yielded tails. To find the average of $f(s)$ over all of these $N_{\text{tot}} = N_1 + N_2$ trials, we sum all the f values and divide by N_{tot}. Equivalently, however, we can rearrange the sum by first adding up all N_1 trials with $f(\text{heads}) = +2$, then adding all N_2 trials with $f(\text{tails}) = -1$:

$$\langle f \rangle = (N_1 f(\text{heads}) + N_2 f(\text{tails}))/N_{\text{tot}}.$$

In the limit of large N_{tot}, this expression is equal to $f(\text{heads})\mathcal{P}(\text{heads}) + f(\text{tails})\mathcal{P}(\text{tails})$, which is the same as Equation 3.19. A similar approach proves the formula for any discrete probability distribution.

The left side of Equation 3.19 introduces an abbreviated notation for the expectation.[21] But brevity comes at a price; if we are considering several different distributions—for example, a set of several coins, each with a different value of ξ—then we may need to write something like $\langle f \rangle_\xi$ to distinguish the answers for the different distributions.

Some random systems generate outcomes that are not numbers. For example, if you ask each of your friends to write down a word "at random," then there's no meaning to questions like "What is the average word chosen?" But we have seen that in many cases, the outcome index does have a numerical meaning. As mentioned in Section 3.3.2, we'll usually use the symbol ℓ, not s, for such situations; then it makes sense to discuss the average value of many draws of ℓ itself, sometimes called the **first moment** of $\mathcal{P}(\ell)$. (The word "first" sets the stage for higher moments, which are expectations of higher powers of ℓ.)

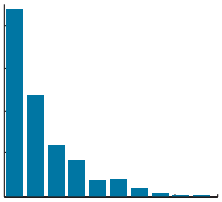


Figure 3.2b (page 38)

Equation 3.19 gives the first moment of a random variable as $\langle \ell \rangle = \sum_\ell \ell \mathcal{P}(\ell)$. Notice that $\langle \ell \rangle$ is a specific number characterizing the distribution, unlike ℓ itself (which is a random value drawn from that distribution), or $\mathcal{P}(\ell)$ (which is a *function* of ℓ). The expectation may not be equal to the **most probable value**, which is the value of ℓ where $\mathcal{P}(\ell)$ attains its maximum.[22] For example, in Figure 3.2b, the most probable value of the waiting time is *zero,* but clearly the average waiting time is greater than that.

Your Turn 3K

Show that $\langle 3 \rangle = 3$. That is, consider a "random variable" whose value on every draw is always exactly equal to 3. More generally, the expectation of any *constant* is simply that constant, regardless of what distribution we use. So, in particular, think about why $\langle (\langle f \rangle) \rangle$ is the same as $\langle f \rangle$.

3.5.2 The variance of a random variable is one measure of its fluctuation

If you measure ℓ just once, you are not guaranteed to observe exactly the most probable value. We use words like "spread," "jitter," "noise," "dispersion," and "fluctuation" to describe

[21] The notations $\langle f \rangle, \mathbb{E}(f), \mu_f$, "expectation of f," "expected value of f," and "expectation value of f" are all synonyms in various cultures for "the mean of an infinitely replicated set of measurements of a random variable." This concept is different from "the mean of a particular, finite set of measurements," which we will call the "sample mean."

[22] The most probable value of a discrete distribution is also called its **mode**. If $\mathcal{P}(\ell)$ attains its maximum value at two or more distinct outcomes, then its most probable value is not defined. A Uniform distribution is an extreme example of this situation.

this phenomenon. It is closely related to the "risk" that Nick and Nora wanted to assess in their coin-toss game. For a Uniform distribution, the "spread" clearly has something to do with how *wide* the range of reasonably likely ℓ values is. Can we make this notion precise, for any kind of distribution?

One way to make these intuitions quantitative is to define[23]

$$\operatorname{var} f = \left\langle \left(f - \langle f \rangle\right)^2 \right\rangle. \quad \textbf{variance of a random variable} \tag{3.20}$$

The right side of Equation 3.20 essentially answers the question, "How much does f deviate from its expectation, on average?" But notice that in this definition, it was crucial to square $(f - \langle f \rangle)$. Had we computed the expectation of $(f - \langle f \rangle)$, we'd have found that the answer was always zero, which doesn't tell us much about the spread of f! By squaring the deviation, we ensure that variations above and below the expectation make reinforcing, not canceling, contributions to the variance.

Like the expectation, the variance depends both on which random variable $f(\ell)$ we are studying and also on the distribution $\mathcal{P}(\ell)$ being considered. Thus, if we study a family of distributions with a parameter, such as ξ for the coin flip, then $\operatorname{var} f$ will be a *function of* ξ. It is not, however, a function of ℓ, because that variable is summed in Equation 3.19.

Another variation on the same idea is the **standard deviation** of f in the given distribution,[24] defined as $\sqrt{\operatorname{var} f}$. The point of taking the square root is to arrive at a quantity with the same dimensions as f.

Example Here's another motivation for introducing the square root into the definition of standard deviation. Imagine a population of Martian students, each exactly twice as tall as a corresponding student in your class. Surely the "spread" of the second distribution should be twice the "spread" of the first. Which descriptor has that property?

Solution The variance for Martian students is $\operatorname{var}(2\ell) = \left\langle \left((2\ell) - \langle 2\ell \rangle\right)^2 \right\rangle = 2^2 \left\langle \left(\ell - \langle \ell \rangle\right)^2 \right\rangle$. Thus, the variance of the Martians' height distribution is *four* times as great as ours. We say that the factor of 2 "inside" the variance became 2^2 when we moved it "outside." The standard deviation, not the variance, scales with a factor of 2.

Example a. Show that $\operatorname{var} f = \langle f^2 \rangle - \left(\langle f \rangle\right)^2$. (If f is ℓ itself, we say, "The variance is the second moment minus the square of the first moment.")
b. Show that, if $\operatorname{var} f = 0$, Equation 3.20 implies that every measurement of f actually does give exactly $\langle f \rangle$.

Solution a. Expand Equation 3.20 to find $\operatorname{var} f = \langle f^2 \rangle - 2\langle f \langle f \rangle \rangle + \left\langle \left(\langle f \rangle\right)^2 \right\rangle$. Now remember that $\langle f \rangle$ is itself a constant, not a random variable. So it can be pulled out of expectations

[23] Section 5.2 will introduce a class of distributions for which the variance is *not* useful as a descriptor of the spread. Nevertheless, the variance is simple, widely used, and appropriate in many cases.
[24] The standard deviation is also called the "root-mean-square" or **RMS deviation** of f. Think about why that's a good name for it.

(see also Your Turn 3K), and we get

$$\operatorname{var} f = \langle f^2 \rangle - 2\big(\langle f \rangle\big)^2 + \big(\langle f \rangle\big)^2,$$

which reduces to what was to be shown.
b. Let $f_* = \langle f \rangle$. We are given that $0 = \langle (f - f_*)^2 \rangle = \sum_\ell \mathcal{P}(\ell)(f(\ell) - f_*)^2$. Every term on the right side is ≥ 0, yet their sum equals zero. So every term is separately zero. For each outcome ℓ, then, we must either have $\mathcal{P}(\ell) = 0$, or else $f(\ell) = f_*$. The outcomes with $\mathcal{P} = 0$ never happen, so every measurement of f yields the value f_*.

Suppose that a discrete random system has outcomes that are labeled by an integer ℓ. We can construct a new random variable m as follows: Every time we are asked to produce a sample of m, we draw a sample of ℓ and add the constant 2. (That is, $m = \ell + 2$.) Then the distribution $\mathcal{P}_{\mathrm{m}}(m)$ has a graph that looks exactly like that of $\mathcal{P}_\ell(\ell)$, but *shifted* to the right by 2, so not surprisingly $\langle m \rangle = \langle \ell \rangle + 2$. Both distributions are equally wide, so (again, not surprisingly) both have the same variance.

Your Turn 3L

a. Prove those two claims, starting from the relevant definitions.
b. Suppose that another random system yields *two* numerical values on every draw, ℓ and s, and the expectations and variances of both are given to us. Find the expectation of $2\ell + 5s$. Express what you found as a general rule for the expectation of a linear combination of random variables.
c. Continuing (b), can you determine the variance of $2\ell + 5s$ from the given information?

Example Find the expectation and variance of the Bernoulli trial distribution, $\mathcal{P}_{\mathrm{bern}}(s; \xi)$, as functions of the parameter ξ.

Solution The answer depends on what numerical values $f(s)$ we assign to heads and tails; suppose these are 1 and 0, respectively. Summing over the sample space just means adding two terms. Hence, $\langle f \rangle = 0 \times (1 - \xi) + 1 \times \xi = \xi$, and

$$\operatorname{var} f = \langle f^2 \rangle - \big(\langle f \rangle\big)^2 = \big(0^2 \times (1 - \xi) + 1^2 \times \xi\big) - \xi^2,$$

or

$$\langle f \rangle = \xi, \qquad \operatorname{var} f = \xi(1 - \xi). \quad \text{for Bernoulli trial} \tag{3.21}$$

Think about why these results are reasonable: The extreme values of ξ (0 and 1) correspond to certainty, or no spread in the results. The trial is most unpredictable when $\xi = \frac{1}{2}$, and that's exactly where the function $\xi(1 - \xi)$ attains its maximum. Try the derivation again, with different values for f(heads) and f(tails).

Your Turn 3M

Suppose that f and g are two independent random variables in a discrete random system.
a. Show that $\langle fg\rangle = \langle f\rangle\langle g\rangle$. Think about how you had to use the assumption of independence and Equation 3.14 (page 49); give a counterexample of two *non*independent random variables that *don't* obey this rule.
b. Find the expectation and variance of $f + g$ in terms of the expectations and variances of f and g separately.
c. Repeat (b) for the quantity $f - g$.
d. Suppose that the expectations of f and g are both greater than zero. Define the **relative standard deviation** (RSD) of a random variable x as $(\text{var}\, x)/|\langle x\rangle|$, a dimensionless quantity. Compare the RSD of $f + g$ with the corresponding quantity for $f - g$.

We can summarize part of what you just found by saying,

*The difference of two noisy variables is a **very** noisy variable.* (3.22)

T_2 *Section 3.5.2′ (page 60) discusses some other moments that are useful as reduced descriptions of a distribution, and some tests for statistical independence of two random variables.*

3.5.3 The standard error of the mean improves with increasing sample size

Suppose that we've got a replicable random system: It allows repeated, independent measurements of a quantity f. We'd like to know the expectation of f, but we don't have time to make an infinite set of measurements; nor do we know a priori the distribution $\mathcal{P}(\ell)$ needed to evaluate Equation 3.19. So we make a finite set of M measurements and average over that, obtaining the **sample mean** $\bar{f}$. This quantity is itself a random variable, because when we make another batch of M measurements and evaluate it, we won't get exactly the same answer.[25] Only in the limit of an infinitely big sample do we expect the sample mean to become a specific number. Because we never measure infinitely big samples in practice, we'd like to know: How good an estimate of the true expectation is $\bar{f}$?

Certainly $\langle\bar{f}\rangle$ is $1/M$ times the sum of M terms, each of which has the same expectation (namely, $\langle f\rangle$). Thus, $\langle\bar{f}\rangle = \langle f\rangle$. But we also need an estimate of how much $\bar{f}$ *varies* from one batch of samples to the next, that is, its variance:

$$\text{var}(\bar{f}) = \text{var}\Big(\frac{1}{M}(f_1 + \cdots + f_M)\Big).$$

Here, f_i is the value that we measured in the ith measurement of a batch. The random variables f_i are all assumed to be independent of one another, because each copy of a replicable system is unaffected by every other one. The constant $1/M$ inside the variance can be replaced by a factor of $1/M^2$ outside.[26] Also, in Your Turn 3M(b), you found that

[25] T_2 More precisely, $\bar{f}$ is a random variable on the joint distribution of batches of M independent measurements.
[26] See page 55.

the variance of the sum of independent variables equals the sum of their variances, which in this case are all equal. So,

$$\mathrm{var}(\bar{f}) = \left(\frac{1}{M^2}M\right)\left(\mathrm{var}\, f\right) = \frac{1}{M}\,\mathrm{var}\, f. \tag{3.23}$$

The factor $1/M$ in this answer means that

The sample mean becomes a better estimate of the true expectation as we average over more measurements. (3.24)

The square root of Equation 3.23 is called the **standard error of the mean**, or **SEM**.

The SEM illustrates a broader idea: A **statistic** is something we compute from a finite sample of data by following a standard recipe. An **estimator** is a statistic that is useful for inferring some property of the underlying distribution of the data. Idea 3.24 says that the sample mean is a useful estimator for the expectation.

THE BIG PICTURE

Living organisms are inference machines, constantly seeking patterns in their world and ways to exploit those regularities. Many of these patterns are veiled by partial randomness. This chapter has begun our study of how to extract whatever discernible, relevant structure can be found from a limited number of observations.

Chapters 4–8 will extend these ideas, but already we have obtained a powerful tool, the Bayes formula (Equation 3.17, page 52). In a strictly mathematical sense, this formula is a trivial consequence of the definition of conditional probability. But we have seen that conditional probability itself is a subtle concept, and one that arises naturally in certain questions that we need to understand (see Section 3.4.3); the Bayes formula clarifies how to apply it.

More broadly, randomness is often a big component of a physical model, and so that model's prediction will in general be a probability distribution. We need to learn how to confront such models with experimental data. Chapter 4 will develop this idea in the context of a historic experiment on bacterial genetics.

KEY FORMULAS

- *Probability distribution of a discrete, replicable random system:* $\mathcal{P}(\ell) = \lim_{N_{\mathrm{tot}}\to\infty} N_\ell/N_{\mathrm{tot}}$. For a finite number of draws N_{tot}, the integers N_ℓ are sometimes called the frequencies of the various possible outcomes; the numbers N_ℓ/N_{tot} all lie between 0 and 1 and can be used as estimates of $\mathcal{P}(\ell)$.
- *Normalization of discrete distribution:* $\sum_\ell \mathcal{P}(\ell) = 1$.
- *Bernoulli trial:* $\mathcal{P}_{\mathrm{bern}}(\underline{\text{heads}};\xi) = \xi$ and $\mathcal{P}_{\mathrm{bern}}(\underline{\text{tails}};\xi) = 1-\xi$. The parameter ξ, and $\mathcal{P}$ itself, are dimensionless. If heads and tails are assigned numerical values $s = 1$ and 0, respectively, then the expectation of the random variable is $\langle s\rangle = \xi$ and the variance is $\mathrm{var}\, s = \xi(1-\xi)$.
- *Addition rule:* $\mathcal{P}(\mathsf{E}_1 \textbf{ or } \mathsf{E}_2) = \mathcal{P}(\mathsf{E}_1) + \mathcal{P}(\mathsf{E}_2) - \mathcal{P}(\mathsf{E}_1 \textbf{ and } \mathsf{E}_2)$.
- *Negation rule:* $\mathcal{P}(\textbf{not-}\mathsf{E}) = 1 - \mathcal{P}(\mathsf{E})$.
- *Product rule:* $\mathcal{P}(\mathsf{E}_1 \textbf{ and } \mathsf{E}_2) = \mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2) \times \mathcal{P}(\mathsf{E}_2)$. (This formula is actually the definition of the conditional probability.)

- *Independence:* Two events are statistically independent if $\mathcal{P}(\mathsf{E}\ \textbf{and}\ \mathsf{E}') = \mathcal{P}(\mathsf{E}) \times \mathcal{P}(\mathsf{E}')$, or equivalently $\mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2) = \mathcal{P}(\mathsf{E}_1 \mid \textbf{not-}\mathsf{E}_2) = \mathcal{P}(\mathsf{E}_1)$.
- *Geometric distribution:* $\mathcal{P}_{\text{geom}}(j;\xi) = \xi(1-\xi)^{(j-1)}$ for discrete, independent attempts with probability of "success" equal to ξ on any trial. The probability $\mathcal{P}$, the random variable $j = 1, 2, \ldots$, and the parameter ξ are all dimensionless. The expectation of j is $1/\xi$, and the variance is $(1-\xi)/(\xi^2)$.
- *Marginal distribution:* For a joint distribution $\mathcal{P}(\ell, s)$, the marginal distributions are $\mathcal{P}_\ell(\ell_0) = \sum_s \mathcal{P}(\ell_0, s)$ and $\mathcal{P}_s(s_0) = \sum_\ell \mathcal{P}(\ell, s_0)$. If ℓ and s are independent, then $\mathcal{P}(\ell|s) = \mathcal{P}_\ell(\ell)$ and conversely; equivalently, $\mathcal{P}(\ell, s) = \mathcal{P}_\ell(\ell)\mathcal{P}_s(s)$ in this case.
- *Bayes:* $\mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2) = \mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_1)\mathcal{P}(\mathsf{E}_1)/\mathcal{P}(\mathsf{E}_2)$. In the context of inferring a model, we call $\mathcal{P}(\mathsf{E}_1)$ the prior distribution, $\mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2)$ the posterior distribution in the light of new information E_2, and $\mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_1)$ the likelihood function.
  Sometimes the formula can usefully be rewritten by expressing the denominator as $\mathcal{P}(\mathsf{E}_2) = \mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_1)\mathcal{P}(\mathsf{E}_1) + \mathcal{P}(\mathsf{E}_2 \mid \textbf{not–}\mathsf{E}_1)\mathcal{P}(\textbf{not–}\mathsf{E}_1)$.
- *Moments:* The expectation of a discrete random variable f is its first moment: $\langle f \rangle = \sum_\ell f(\ell)\mathcal{P}(\ell)$. The variance is the mean-square deviation from the expected value: $\text{var}\,\ell = \langle (\ell - \langle \ell \rangle)^2 \rangle$. Equivalently, $\text{var}\,\ell = \langle \ell^2 \rangle - (\langle \ell \rangle)^2$. The standard deviation is the square root of the variance. T2 Skewness and kurtosis are defined in Section 3.5.2′ (page 60).
- T2 *Correlation and covariance:* $\text{cov}(\ell, s) = \langle (\ell - \langle \ell \rangle)(s - \langle s \rangle) \rangle$.
  $\text{corr}(\ell, s) = \text{cov}(\ell, s)/\sqrt{(\text{var}\,\ell)(\text{var}\,s)}$.

FURTHER READING

Semipopular:
Conditional probability and the Bayes formula: Gigerenzer, 2002; Mlodinow, 2008; Strogatz, 2012; Woolfson, 2012.

Intermediate:
Bolker, 2008; Denny & Gaines, 2000; Dill & Bromberg, 2010, chapt. 1; Otto & Day, 2007, §P3.

Technical:
Gelman et al., 2014.

T_2 **Track 2**

3.4.1′a Extended negation rule

Here is another useful fact about conditional probabilities:

Your Turn 3N

a. Show that $\mathcal{P}(\textbf{not-}\mathsf{E}_1 \mid \mathsf{E}_2) = 1 - \mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2)$.
b. More generally, find a normalization rule for $\mathcal{P}(\ell \mid \mathsf{E})$, where ℓ is a discrete random variable and E is any event.

3.4.1′b Extended product rule

Similarly,

$$\mathcal{P}(\mathsf{E}_1 \textbf{ and } \mathsf{E}_2 \mid \mathsf{E}_3) = \mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2 \textbf{ and } \mathsf{E}_3) \times \mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_3). \tag{3.25}$$

Your Turn 3O

Prove Equation 3.25.

3.4.1′c Extended independence property

We can extend the discussion in Section 3.4.1 by saying that E_1 and E_2 are "independent under condition E_3" if knowing E_2 gives us no additional information about E_1 beyond what we already had from E_3; that is,

$$\mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2 \textbf{ and } \mathsf{E}_3) = \mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_3). \qquad \text{independence under condition } \mathsf{E}_3$$

Substituting into Equation 3.25 shows that, if two events are independent under a third condition, then

$$\mathcal{P}(\mathsf{E}_1 \textbf{ and } \mathsf{E}_2 \mid \mathsf{E}_3) = \mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_3) \times \mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_3). \tag{3.26}$$

T_2 **Track 2**

3.4.4′ Generalized Bayes formula

There is a useful extension of the Bayes formula that states

$$\mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_2 \textbf{ and } \mathsf{E}_3) = \mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_1 \textbf{ and } \mathsf{E}_3) \times \mathcal{P}(\mathsf{E}_1 \mid \mathsf{E}_3) / \mathcal{P}(\mathsf{E}_2 \mid \mathsf{E}_3). \tag{3.27}$$

Your Turn 3P

Use your result in Your Turn 3O to prove Equation 3.27.

T_2 **Track 2**

3.5.2′a Skewness and kurtosis

The first and second moments of a distribution, related to the location and width of its peak, are useful summary statistics, particularly when we repackage them as

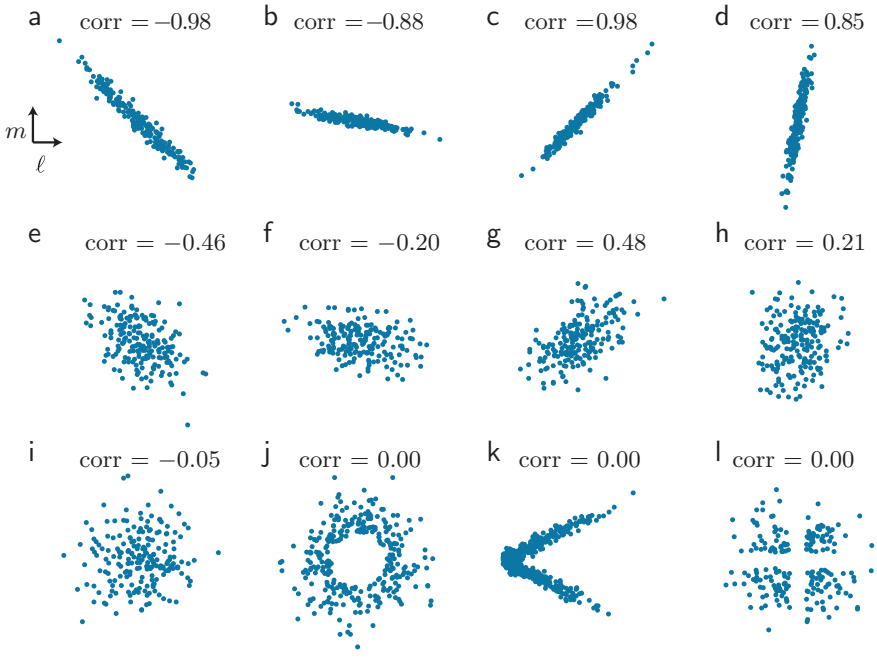


Figure 3.6 [Simulated datasets.] **Correlation coefficients of some distributions.** Each panel shows a cloud representation of a joint probability distribution, as a set of points in the ℓ-m plane; the corresponding value for corr(ℓ, m) is given above each set. Note that the correlation coefficient reflects the noisiness and direction of a linear relationship (a–h), and it's zero for independent variables (i), but it misses other kinds of correlation (j–l). In each case, the correlation coefficient was estimated from a sample of 5000 points, but only the first 200 are shown.

expectation and variance. Two other moments are often used to give more detailed information:

- Some distributions are asymmetric about their peak. The asymmetry can be quantified by computing the **skewness**, defined by $\left\langle \left(\ell - \langle \ell \rangle\right)^3 \right\rangle / (\text{var}\,\ell)^{3/2}$. This quantity equals zero for any symmetric distribution.
- Even if two distributions each have a single, symmetric peak, and both have the same variance, nevertheless their peaks may not have the same shape. The **kurtosis** further specifies the peak shape; it is defined as $\left\langle \left(\ell - \langle \ell \rangle\right)^4 \right\rangle / (\text{var}\,\ell)^2$.

3.5.2′b Correlation and covariance

The product rule for independent events (Equation 3.12) can also be regarded as a *test* for whether two events are statistically independent. This criterion, however, is not always easy to evaluate. How can we tell from a joint probability distribution $\mathcal{P}(\ell, m)$ whether it can be written as a product? One way would be to evaluate the conditional probability $\mathcal{P}(\ell \mid m)$ and see, for every value of ℓ, whether it depends on m. But there is a shortcut that can at least show that two variables are *not* independent (that is, that they are correlated).

Suppose that ℓ and m both have numerical values; that is, both are random variables. Then we can define the **correlation coefficient** as an expectation:

$$\operatorname{corr}(\ell, m) = \frac{\big\langle \big(\ell - \langle \ell \rangle\big)\big(m - \langle m \rangle\big)\big\rangle}{\sqrt{(\operatorname{var} \ell)(\operatorname{var} m)}}. \tag{3.28}$$

Your Turn 3Q

a. Show that the numerator in Equation 3.28 may be replaced by $\langle \ell m \rangle - \langle \ell \rangle \langle m \rangle$ without changing the result. [*Hint:* Go back to the Example on page 55 concerning variance.]
b. Show that $\operatorname{corr}(\ell, m) = 0$ if ℓ and m are statistically independent.
c. Explain why it was important to subtract the expectation from each factor in parentheses in Equation 3.28.

The numerator of Equation 3.28 is also called the **covariance** of ℓ and m, or $\operatorname{cov}(\ell, m)$. Dividing by the denominator makes the expression independent of the overall scale of ℓ and m; this makes the value of the correlation coefficient a meaningful descriptor of the tendency of the two variables to track each other.

The correlation coefficient gets positive contributions from every measurement in which ℓ and m are both larger than their respective expectations, but also from every measurement in which both are *smaller* than their expectations. Thus, a positive value of $\operatorname{corr}(\ell, m)$ indicates a roughly linear, increasing relationship (Figure 3.6c,d,g,h). A negative value has the opposite interpretation (Figure 3.6a,b,e,f).

When we flip a coin repeatedly, there's a natural linear ordering according to time: Our data form a **time series**. We don't expect the probability of flipping heads on trial i to depend on the results of the previous trials, and certainly not on those of future trials. But many other time series do have such dependences.[27] To spot them, assign numerical values $f(s)$ to each flip outcome and consider each flip $f_1, \ldots, f_M$ in the series to be a different random variable, in which the f_i's may or may not be independent. If the random system is stationary (all probabilities are unchanged if we shift every index by the same amount), then we can define its **autocorrelation function** as $C(j) = \operatorname{cov}(f_i, f_{i+j})$ for any starting point i. If this function is nonzero for any j (other than $j = 0$), then the time series is correlated.

3.5.2′c Limitations of the correlation coefficient

Equation 3.28 introduced a quantity that equals zero if two random variables are statistically independent. It follows that if the correlation coefficient of two random variables is nonzero, then they are correlated. However, the converse statement is not always true: It is possible for two nonindependent random variables to have correlation coefficient equal to zero. Panels (j–l) of Figure 3.6 show some examples. For example, panel (j) represents a distribution with the property that ℓ and m are never both close to zero; thus, knowing the value of one tells something about the value of the other, even though there is no linear relation.

[27] For example, the successive positions of a particle undergoing Brownian motion (example **5** on page 38).

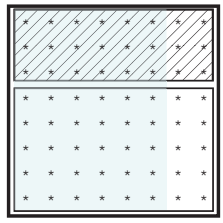
PROBLEMS

3.1 Complex time series

Think about the weather—for example, the daily peak temperature. It's proverbially unpredictable. Nevertheless, there are several kinds of structure to this time series. Name a few and discuss.

3.2 Medical test

Look at Figures 3.4a,b. If E_2 is the outcome of a medical test and E_1 is the statement that the patient is actually sick, then which of these figures describes a better test?



Figures 3.4a (page 46)

3.3 Six flips

Write a few lines of computer code to make distributions like Figures 3.1a,b, but with $m = 6$ and 6000 total draws. (That is, generate 6000 six-bit random binary fractions.) If you don't like what you see, explain and then fix it.

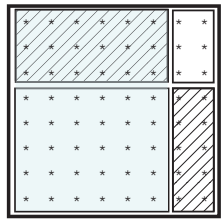


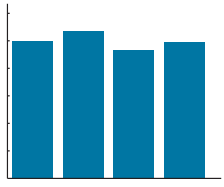
Figure 3.4b (page 46)

3.4 Random walk end point distribution

This problem introduces the "random walk," a physical model for Brownian motion. Get Dataset 4, which contains experimental data. The two columns represent the x and y coordinates of the displacements of a particle undergoing Brownian motion, observed at periodic time intervals (see Figure 3.3a,b).

a. Tabulate the values of $x^2 + y^2$ in the experimental data, and display them as a histogram.

Suppose that a chess piece is placed on a line, initially at a point labeled 0. Once per second, the chess piece is moved a distance $d = 1\ \mu$m along either the + or − direction. The choice is random, each direction is equally probable, and each step is statistically independent of all the others. We imagine making many trajectories, all starting at 0.



Figures 3.1a (page 37)

b. Simulate 1000 two-dimensional random walks, all starting at the origin. To do this, at each step randomly choose $\Delta x = \pm 1\ \mu$m and also $\Delta y = \pm 1\ \mu$m. Display the histogram of $x^2 + y^2$ after 500 steps, and compare your answer qualitatively with the experimental result.

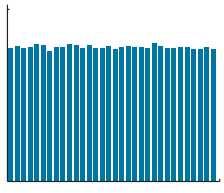


Figure 3.1b (page 37)

c. Do your results suggest a possible mathematical form for this distribution? How could you replot your data to check this idea?

3.5 Gambler's fallacy

There seems to be a hardwired misperception in the human brain that says, "If I've flipped heads five times in a row, that increases the probability that I'll get tails the next time." Intellectually, we know that's false, but it's still hard to avoid disguised versions of this error in our daily lives.

If we call heads +1 and tails −1, we can let X be the sum of, say, 10 flips. On any given 10-flip round we probably won't get exactly zero. But if we keep doing 10-flip rounds, then the long-term average of X is zero.

Suppose that one trial starts out with five heads in a row. We wish to check the proposition

> *"My next five flips will be more than half tails, in order to pull X closer to zero, because X 'wants' to be zero on average."*

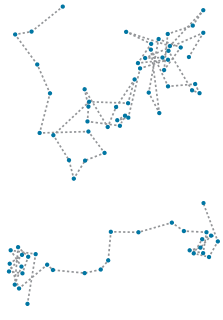


Figure 3.3a,b (page 39)

a. Use a computer to simulate 200 ten-flip sequences, pull out those few that start with five heads in a row, and find the average value of X among only those few sequences.

[*Hint:* Define variables called `Ntrials` and `Nflips`; use those names throughout your code. At the top, say `Ntrials=200, Nflips=10`.]

b. Repeat (a) with `Ntrials` = 2000 and 8000 sequences. Does the answer seem to be converging to zero, as predicted in the quoted text above? Whatever your answer, give some explanation for why the answer you found should have been expected.

c. To understand what "regression to the mean" actually means, repeat (a) but with 50 000 sequences of `Nflips` flips. Consider `Nflips` = 10, 100, 500, and 2500. [*Hint:* Instead of writing similar code four times, write it once, but put it inside a loop that gives `Nflips` a new value each time it runs.]

d. As you get longer and longer sequences (larger values of `Nflips`), your answer in (c) will become insignificant compared with the *spread* in the results among trials. Confirm this as follows. Again, start with `Nflips` = 10. For each of your sequences, save the value of X, creating a list with 50 000 entries. Then find the spread (standard deviation) of all the values you found. Repeat with `Nflips` = 100, 500, and 2500. Discuss whether the proposition

> *The effect of unusual past behavior doesn't disappear; it just gets* ***diluted*** *as time goes on.*

is more appropriate than the idea in quotes above.

3.6 Virus evolution

The genome of the HIV virus, like any genome, is a string of "letters" (base pairs) in an "alphabet" containing only four letters. The message for HIV is rather short, just $n \approx 10^4$ letters in all.

The probability of errors in reverse transcribing the HIV genome is about one error for every $3 \cdot 10^4$ "letters" copied. Suppose that each error replaces a DNA base by one of the three other bases, chosen at random. Each time a virion infects a T cell, the reverse transcription step creates opportunities for such errors, which will then be passed on to the offspring virions. The total population of infected T cells in a patient's blood, in the quasi-steady state, is roughly 10^7 (see Problem 1.7).

a. Find the probability that a T cell infection event will generate one *particular* error, for example, the one lucky spontaneous mutation that could confer resistance to a drug. Multiply by the population to estimate the number of T cells already present with a specified mutation, prior to administering any drug. Those cells will later release resistant virions.

b. Repeat (a), but this time for the probability of spontaneously finding *two* or *three* specific errors, and comment.

[*Note:* Make the conservative approximation that each infected T cell was infected by a wild-type virion, so that mutations do not accumulate. For example, the wild-type may reproduce faster than the mutant, crowding it out in the quasi-steady state.]

3.7 Weather

Figure 3.7a is a graphical depiction of the probability distribution for the weather on consecutive days in an imagined place and season. The outcomes are labeled X_1X_2, where X = r or s indicates the mutually exclusive options "rain" or "sunny," and the subscripts 1 and 2 denote today and tomorrow, respectively.

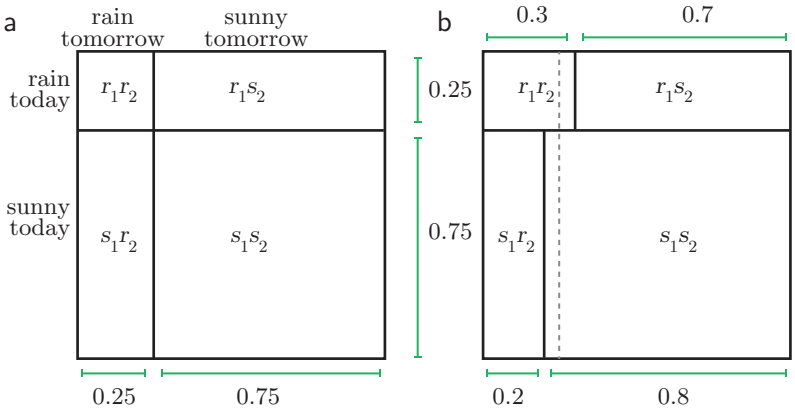


Figure 3.7 [Box diagrams.] **Probabilities for outcomes on consecutive days.** (a) A case where the two outcomes are independent. (b) A modified set of probabilities. The *dashed line* indicates the situation in (a) for comparison.

Panel (b) shows a more realistic situation. Compute $\mathcal{P}(\text{rain tomorrow})$, $\mathcal{P}(\text{rain tomorrow} \mid \text{rain today})$, and $\mathcal{P}(\text{rain tomorrow} \mid \text{sunny today})$ and comment. Repeat for the situation in panel (a).

3.8 Family history

Review Section 3.4.3. How does the probability of being sick given a positive test result change if you also know that you have some family history predisposing you to the disease? Discuss how to account for this information by using the Bayes formula.

3.9 Doping in sports

Background: A laboratory flagged a cyclist based on a urine sample taken following stage 17 of the 2006 Tour de France. The lab claimed that the test was highly unlikely to turn out positive unless the subject had taken illegal steroid drugs. Based on this determination, the International Court of Arbitration for Sport upheld doping charges against the cyclist.

In fact, the cyclist was tested eight times during the race, and a total of 126 tests were made on all contestants.

a. Suppose that the cyclist was innocent, but the false-positive rate of the test was 2%. What is the probability that at least one of the 8 tests would come out positive?

b. Suppose that the false-positive rate was just 1% and that *all* contestants in the race were innocent. What is the chance that *some* contestant (that is, one or more) would test positive at least once?

c. Actually, it's not enough to know the false-positive rate. If we wish to know the probability of guilt given the test results, we need one *additional* piece of quantitative information (which the court did not have). What is that needed quantity? [*Hint:* You may assume that the false-negative rate is small. It is not the quantity being requested.]

3.10 Hemoccult test

Figure 3.5 (page 51) represents a situation in which the sensitivity of a medical test is equal to its selectivity. This is actually not a very common situation.

The hemoccult test, among others, is used to detect colorectal cancer. Imagine that you conduct mass screening with this test over a certain region of the country, in a particular age group. Suppose that, in the absence of any other information, 0.3% of individuals in this group are known to have this disease. People who have the disease are 50% likely to have a positive test. Among those who do not have the disease, 3% nevertheless test positive.

Suppose that a randomly chosen participant tests positive. Based on the above data, and that single test, what can you say about $\mathcal{P}(\text{sick} \mid \text{pos})$?

3.11 Smoking and cancer

In 1993, about 28% of American males were classified as cigarette smokers. The probability for a smoker to die of lung cancer in a given period of time was about 11 times the probability for a nonsmoker to die of lung cancer in that period.

a. Translate these statements into facts about $\mathcal{P}(\text{die of lung cancer} \mid \text{smoker})$, $\mathcal{P}(\text{die of lung cancer} \mid \text{nonsmoker})$, $\mathcal{P}(\text{smoker})$, and $\mathcal{P}(\text{nonsmoker})$.

b. From these data, compute the probability that an American male who died of lung cancer in the specified period was a smoker.

3.12 Effect of new information

The "Monty Hall" puzzle is a classic problem that can be stated and analyzed in the language we are developing.

A valuable prize is known to lie behind one of three closed doors. All three options are equally probable. The director of the game ("Monty") knows which door conceals the prize, but you don't. The rules state that after you make a preliminary choice, Monty will choose one of the *other* two doors, open it, and reveal that the prize is *not* there. He then gives you the option of changing your preliminary choice, or sticking with it. After you make this decision, your final choice of door is opened. The puzzle is to find the best strategy for playing this game.

Let's suppose that you initially choose door #1.[28] Certainly, either #2 or #3, or both, has no prize. After Monty opens one of these doors, have you now gained any relevant additional information? If not, there's no point in changing your choice (analogous to scenario **a** in Section 3.4.1). If so, then maybe you should change (analogous to scenario **b**). To analyze the game, make a grid with six cells:

		It's actually behind door #		
		1	2	3
Monty reveals it's not behind door #	2	*A*	*B*	*C*
	3	*D*	*E*	*F*

In this table,

$$A = \mathcal{P}\big(\text{it's behind door \#1 and Monty shows you it's not behind \#2}\big),$$

$$D = \mathcal{P}\big(\text{it's behind door \#1 and Monty shows you it's not behind \#3}\big),$$

[28] By symmetry, it's enough to analyze only this case.

and so on. Convince yourself that

$A = 1/6,\ D = 1/6$, but

$B = 0,\ C = 1/3$ (Monty has no choice if it's not behind the door you chose), and

$E = 1/3,\ F = 0$.

a. Now compute $\mathcal{P}\big(\text{it's behind \#1} \,|\, \text{Monty showed you \#2}\big)$ by using the definition of conditional probability.
b. Compute $\mathcal{P}\big(\text{it's behind \#3} \mid \text{Monty showed you \#2}\big)$, and compare it with your answer in (a). Also compute $\mathcal{P}\big(\text{it's behind \#2} \mid \text{Monty showed you \#2}\big)$. (The second quantity is zero, because Monty wouldn't do that.)
c. Now answer this question: If you initially chose #1, and then Monty showed you #2, should you switch your initial choice to #3 or remain with #1?

3.13 Negation rule

a. Suppose that you are looking for a special type of cell, perhaps those tagged by expressing a fluorescent protein. You spread a drop of blood on a slide marked with a grid containing N boxes, and examine each box for the cell type of interest. Suppose that a particular sample has a total of M tagged cells. What is the probability that at least one box on the grid contains more than one of these M cells? [*Hint:* Each tagged cell independently "chooses" a box, so each has probability $1/N$ to be in any particular box. Use the product rule to compute the probability that *no* box on the grid has more than one tagged cell, and then use the negation rule.]
b. Evaluate your answer for $N = 400$, $M = 20$.

3.14 Modified Bernoulli trial

The Example on page 56 found the expectation and variance of the Bernoulli trial distribution as functions of its parameter ξ, if heads is assigned the numerical value 1 and tails 0. Repeat, but this time, heads counts as $1/2$ and tails as $-1/2$.

3.15 Perfectly random?

Let ℓ be an integer random variable with the Uniform distribution on the range $3 \le \ell \le 6$. Find the variance of ℓ.

3.16 T2 Variance of a general sum

In Your Turn 3M(b) (page 57), you found the variance of the sum of two random variables, assuming that they were independent. Generalize your result to find the variance of $f + g$ in terms of var f, var g, and the covariance $\text{cov}(f, g)$, *without* assuming independence.

3.17 T2 Multiple tests

Suppose that you are a physician. You examine a patient, and you think it's quite likely that she has strep throat. Specifically, you believe this patient's symptoms put her in a group of people with similar symptoms, of whom 90% are sick. But now you refine your estimate by taking throat swabs and sending them to a lab for testing.

The throat swab is not a perfect test. Suppose that if a patient is sick with strep, then in 70% of cases, the test comes back positive; the rest of the time, it's a false negative. Suppose that, if a patient is not sick, then in 90% of cases, the test comes back negative; the rest of the time, it's a false positive.

You run five successive swabs from the same patient and send them to the lab, where they are all tested independently. The results come back $(+\ -\ +\ -\ +)$, apparently a total muddle. You'd like to know whether any conclusion can be drawn from such data. Specifically, do they revise your estimate of the probability that the patient is sick?

a. Based on this information, what is your new estimate of the probability that the patient is sick? [*Hint:* Prove, then use, the result about independence stated in Equations 3.25–3.26 on page 60.]

b. Work the problem again, but this time from the viewpoint of a worker at the lab, who has no information about the patient other than the five test results. This worker interprets the information in the light of a prior assumption that the patient's chance of being sick is 50% (not 90%).

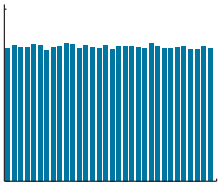
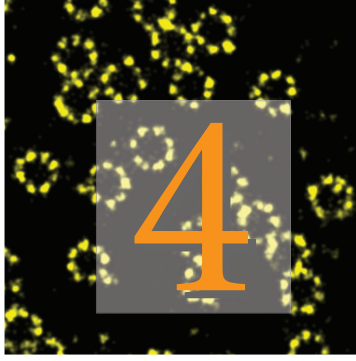


Figure 3.1b (page 37)

3.18 T_2 Binary fractions

Find the expectation and variance of the random, m-bit binary fractions discussed in Section 3.2.1 on page 36 (see Figure 3.1). Use an analytic (exact) argument, not a computer simulation.



Some Useful Discrete Distributions

It may be that universal history is the history of the different intonations given a handful of metaphors.
—Jorge Luis Borges

4.1 Signpost

Much of the everyday business of science involves proposing a model for some phenomenon of interest, poking the model until it yields some quantitative prediction, and then testing the prediction. A theme of this book is that often what is predicted is a probability distribution. This chapter begins our discussion of how to make such predictions, starting from a proposed physical model of a living system.

Chapter 3 may have given the impression that a probability distribution is a purely empirical construction, to be deduced from repeated measurements (via Equation 3.3, page 42). In practice, however, we generally work with distributions that embody simplifying hypotheses about the system (the physical model). For example, we may have reason to believe that a variable is Uniformly distributed on some range. Generally we need more complicated distributions than that, but perhaps surprisingly, just *three* additional discrete distributions describe many problems that arise in biology and physics: the Binomial, Poisson, and Geometric distributions. We'll see that, remarkably, all three are descendants of the humble Bernoulli trial.[1] Moreover, each has rather simple mathematical properties. Knowing some general facts about a distribution at once gives useful information about all the systems to which it applies.

[1] Later chapters will show that the Gaussian and Exponential distributions, and the Poisson process, are also offshoots of Bernoulli.

Our Focus Question is
Biological question: How do bacteria become resistant to a drug or virus that they've never encountered?
Physical idea: The Luria-Delbrück experiment tested a model by checking a statistical prediction.

4.2 Binomial Distribution

4.2.1 Drawing a sample from solution can be modeled in terms of Bernoulli trials

Here is a question that arises in the lab: Suppose that you have 10 mL of solution containing just *four molecules* of a particular type, each of which is tagged with a fluorescent dye. You mix well and withdraw a 1 mL sample (an "aliquot"). How many of those four molecules will be in your sample?[2] One reply is, "I can't predict that; it's random," and of course that is true. But the preceding chapter suggested some more informative questions we can ask about this system.

What we really want to know is a *probability distribution* for the various values for ℓ, the number of molecules in the sample. To determine that distribution, we imagine preparing many identical solutions, extracting a 1 mL sample from each one, and counting how many labeled molecules are in each such sample. Prior to sampling, each labeled molecule wanders at random through the solution, independently of the others. At the moment of sampling, each molecule is captured or not, in a Bernoulli trial with probability ξ. Assigning the value $s = 1$ to capture and 0 to noncapture, we have that $\ell = s_1 + \cdots + s_M$, where M is the total number of tagged molecules in the original solution.

The Bernoulli trial is easy to characterize. Its probability distribution is just a graph with two bars, of heights ξ and $\xi' = 1 - \xi$. If either ξ or ξ' equals 1, then there's no randomness; the "spread" is zero. If $\xi = \xi' = \frac{1}{2}$, the "spread" is maximal (see the Example on page 56). For the problem at hand, however, we have batches of *several* Bernoulli trials (M of them in a batch). We are interested only in a reduced description of the outcomes, not the details of every individual draw in a batch. Specifically, we want the distribution, across batches, for the discrete random variable ℓ.

Before proceeding, we should first try to frame some expectations. The capture of each labeled molecule is like a coin flip. If we flip a fair coin 50 times, we'd expect to get "about" 25 heads, though we wouldn't be surprised to get 24 or 26.[3] In other words, we expect for a fair coin that the most probable value of ℓ is $M/2$; but we also expect to find a spread about that value. Similarly, when we draw an aliquot from solution, we expect to get about ξM tagged molecules in each sample, with some spread.

For *10 000* coin flips, we expect the fraction coming up heads to equal 1/2 to high accuracy, whereas for just a few flips we're not surprised at all to find some extreme results, even $\ell = 0$ or $\ell = M$. For a general Bernoulli trial, we expect the actual number not to deviate much from ξM, if that number is large. Let's make these qualitative hunches more precise.

---

[2] Modern biophysical methods really can give exact counts of individual fluorescent dye molecules in small volumes, so this is not an academic example.
[3] In fact, if we got exactly 25 heads, and redid the whole experiment many times and *always* got exactly 25, *that* would be surprising.

4.2.2 The sum of several Bernoulli trials follows a Binomial distribution

Sampling from solution is like flipping M coins, but recording only the *total* number ℓ of heads that come up. Thus, an "outcome" is one of the aggregate values $\ell = 0, \ldots, M$ that may arise. We'd like to know the probability of each outcome.

The problem discussed in Section 4.2.1 had $M = 4$, and

$$\xi = (\text{sample volume})/(\text{total volume}) = 0.1.$$

If we define $\xi' = 1 - \xi$, then certainly $(\xi + \xi')^4 = 1$. To see why this fact is useful, expand it, to get 16 terms that are guaranteed to add up to 1. Collecting the terms according to powers of ξ and ξ', we find one term containing ξ^4, four terms containing $\xi^3\xi'$, and so on. Generally, the term $\xi^\ell(\xi')^{M-\ell}$ corresponds to flipping heads exactly ℓ times, and by the binomial theorem it contributes

$$\mathcal{P}_{\text{binom}}(\ell; \xi, M) = \frac{M!}{\ell!(M-\ell)!}\,\xi^\ell(1-\xi)^{M-\ell} \text{ for } \ell = 0, \ldots, M \qquad \textbf{Binomial distribution} \tag{4.1}$$

to the total probability (see Figure 4.1). This probability distribution is really a *family* of discrete distributions of ℓ, with two parameters M and ξ. By its construction, it has the normalization property: We get 1 when we sum it over ℓ, holding the two parameters fixed.

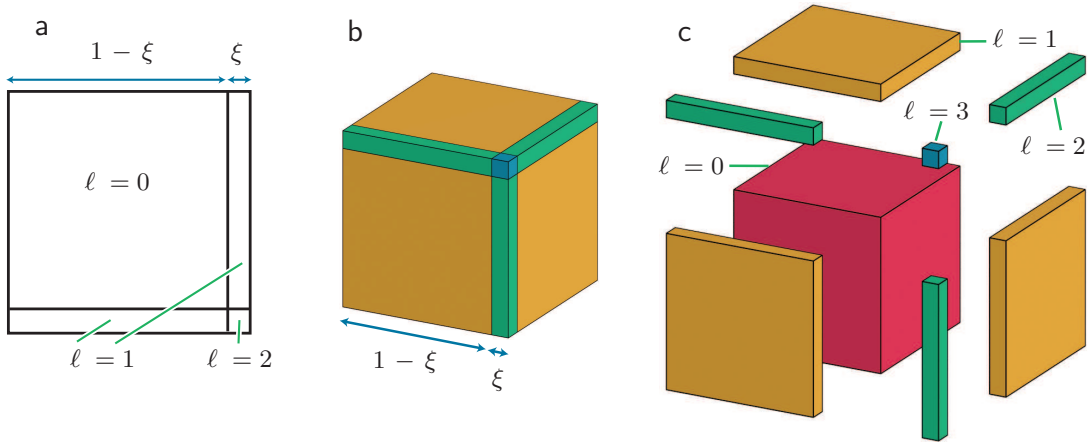


Figure 4.1 [Sketches.] **Graphical representation of the binomial theorem.** (a) For $M = 2$ and $\xi = 1/10$, the small block representing two heads has area ξ^2; the two blocks representing one heads/one tails have combined area $2\xi(1-\xi)$, and the remaining block has area $(1-\xi)^2$. Thus, the three classes of outcomes have areas corresponding to the expressions in Equation 4.1 with $M = 2$ and $\ell = 0$, 1, and 2. The sum of these expressions equals the area of the complete unit square, so the distribution is properly normalized. (b) For $M = 3$, the small cube in the front represents all three flips coming up heads, and so on. (The large cube representing $\ell = 0$ is hidden in the back of the picture.) This time there are four classes of outcomes, again with volumes that correspond to terms of Equation 4.1. (c) Exploded view of panel (b).

Your Turn 4A

a. Evaluate the Binomial distribution for $M = 4$ and $\xi = 0.1$. Is there any significant chance of capturing more than one tagged molecule?
b. Expand the $M = 5$ case, find all six terms, and compare them with the values of $\mathcal{P}_{\text{binom}}(\ell; \xi, M)$ in the general formula above.

4.2.3 Expectation and variance

Example a. What are the expectation and variance of ℓ in the Binomial distribution? [*Hint:* Use the Example on page 56.]
b. Use your answer to (a) to confirm and make precise our earlier intuition that we should get about $M\xi$ heads, and that for large M we should get very little spread about that value.

Solution a. The expectation is $M\xi$, and the variance is $M\xi(1 - \xi)$. These are very easy when we recall the general formulas for expectation and variance for the sum of independent random variables.[4]
b. More precisely, we'd like to see whether the standard deviation is small relative to the expectation. Indeed, their ratio is $\sqrt{M\xi(1-\xi)}/(M\xi)$, which gets small for large enough M.

4.2.4 How to count the number of fluorescent molecules in a cell

Some key molecular actors in cells are present in small numbers, perhaps a few dozen copies per cell. We are often interested in measuring that number as exactly as possible, throughout the life of the cell.

Later chapters will discuss methods that allow us to visualize specific molecules, by making them glow (fluoresce). We'll see that in some favorable cases, it may be possible to see such fluorescent molecules individually, and so to count them directly. In other situations, the molecules move too fast, or otherwise do not allow direct counting. Even then, however, we do know that the molecules are all identical, so their total light output (fluorescence intensity), y, equals their number M times some constant α. Why not just measure y as a proxy for M?

The problem is that it is hard to estimate accurately the constant of proportionality, α, needed to convert the observable y into the desired quantity M. This constant depends on how brightly each molecule fluoresces, how much of its light is lost between emission and detection, and so on. N. Rosenfeld and coauthors found a method to *measure* α, by using a probabilistic argument. They noticed that cell division in bacteria divides the cell's volume into very nearly equal halves. If we know that just prior to division there are M_0 fluorescent molecules, then after division one daughter cell gets M_1 and the other gets $M_2 = M_0 - M_1$. If, moreover, the molecules wander at random inside the cell, then for given M_0 the quantity M_1 will be distributed according to $\mathcal{P}_{\text{binom}}(M_1; M_0, 1/2)$. Hence, the variance of M_1 equals $\frac{1}{2}(1 - \frac{1}{2})M_0$. Defining the "error of partitioning" $\Delta M = M_1 - M_2$ then gives $\Delta M = M_1 - (M_0 - M_1) = 2M_1 - M_0$.

[4] See Your Turn 3M (page 57).

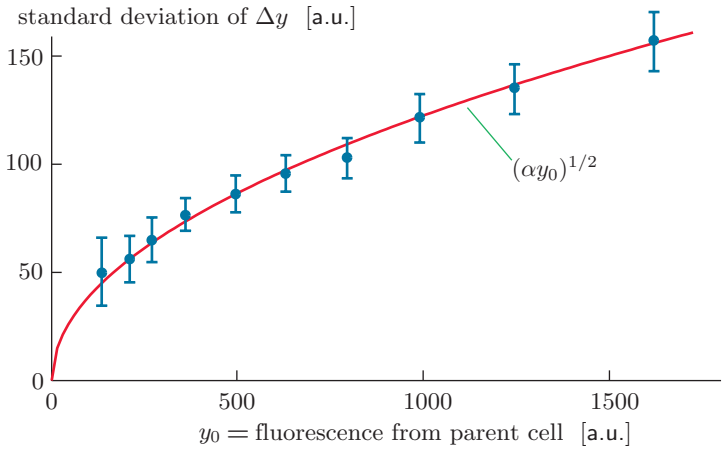


Figure 4.2 [Experimental data with fit.] **Calibration of a single-molecule fluorescence measurement.** *Horizontal axis:* Measured fluorescence intensity of cells prior to division. *Vertical axis:* Sample standard deviation of the partitioning error of cell fluorescence after division. *Error bars* indicate that this quantity is uncertain due in part to the finite number of cells observed. *Red curve:* The predicted function from Idea 4.2. The best-fit value of the parameter α is 15 fluorescence units per tagged molecule. [Data from Rosenfeld et al., 2005.]

Thus,[5]

$$\mathrm{var}(\Delta M) = 4\,\mathrm{var}(M_1) = M_0.$$

We wish to re-express this result in terms of the observed fluorescence, so let $y = \alpha M$, where α is the constant we are seeking:

$$\mathrm{var}(\Delta y) = \alpha^2\,\mathrm{var}(\Delta M) = \alpha^2 M_0 = \alpha y_0.$$

That is, we have predicted that

> *The standard deviation of Δy, among a population of cells all with the same initial fluorescence y_0, is $(\alpha y_0)^{1/2}$.* (4.2)

Idea 4.2 involves some experimentally measurable quantities (y_0 and Δy), as well as the unknown constant α. Fitting this model to data thus yields the desired value of α. The experimenters observed a large number of cells just prior to and just after division;[6] thus, for each value of y_0 they found many values of Δy. Computing the variance gave them a dataset to fit to the prediction in Idea 4.2. Figure 4.2 shows that the data do give a good fit.

4.2.5 Computer simulation

It is nice to have an exact formula like Equation 4.1 for a probability distribution; sometimes important results can be proved directly from such a formula. Other times, however, a known distribution is merely the starting point for constructing something more elaborate, for which exact results are not so readily available. In such a case, it can be important to

[5] See Your Turn 3L(a) (page 56).

[6] T_2 Rosenfeld and coauthors arranged to have a wide range of y_0 values, and they ensured that the fluorescent molecule under study was neither created nor significantly cleared during the observed period of cell division.

simulate the distribution under study, that is, to program a computer to emit sequences of random outcomes with some given distribution.[7] Chapter 3 described how to accomplish this for the Bernoulli trial.[8] Your computer math system may also have a built-in function that simulates sampling from the Binomial distribution, but it's valuable to know how to build such a generator from scratch, for *any* discrete distribution.

We wish to extend the idea of Section 3.2.2 to sample spaces with more than two outcomes. Suppose that we wish to simulate a variable ℓ drawn from $\mathcal{P}_{\text{binom}}(\ell; M, \xi)$ with $M = 3$. We do this by partitioning the unit segment into four bins of widths $(1-\xi)^3$, $3\xi(1-\xi)^2$, $3\xi^2(1-\xi)$, and ξ^3, corresponding to $\ell = 0$, 1, 2, and 3 heads, respectively (see Equation 4.1). The first bin thus starts at 0 and ends at $(1-\xi)^3$, and so on.

Your Turn 4B

a. Write a short computer code that sets up a function `binomSimSetup(xi)`. This function should accept a value of ξ and return a *list* of the locations of the bin edges appropriate for computing $\mathcal{P}_{\text{binom}}(\ell; M = 3, \xi)$ for three-flip sequences.
b. Write a short "wrapper" program that calls `binomSimSetup`. The program should then use the list of bin edges to generate 100 Binomial-distributed values of ℓ and histogram them. Show the histogram for a few different values of ξ, including $\xi = 0.6$.
c. Find the sample mean and the variance of your 100 samples, and compare your answers with the results found in the preceding Example. Repeat with 10 000 samples.

4.3 Poisson Distribution

The formula for the Binomial distribution, Equation 4.1, is complicated. For example, it has two parameters, M and ξ. Two may not sound like a large number, but fitting data to a model rapidly becomes complicated and unconvincing when there are too many parameters. Fortunately, often a simpler, approximate form of this distribution can be used instead. The simplified distribution to be derived in this section has just *one* parameter, so using it can improve the predictive power of a model.

The derivation that follows is so fundamental that it's worth following in detail. It's important to understand the approximation we will use, in order to say whether it is justified for a particular problem.

4.3.1 The Binomial distribution becomes simpler in the limit of sampling from an infinite reservoir

Here is a physical question similar to the one that introduced Section 4.2.1, but with more realistic numbers: Suppose that you take a liter of pure water ($10^6\ \text{mm}^3$) and add five million fluorescently tagged molecules. You mix well, then withdraw one cubic millimeter. How many tagged molecules, ℓ, will you get in your sample?

Section 4.2 sharpened this question to one involving a Binomial probability distribution. For the case under study now, the expectation of that distribution is $\langle \ell \rangle = M\xi = (5 \cdot 10^6)(1\ \text{mm}^3)/(10^6\ \text{mm}^3) = 5$. Suppose next that we instead take a cubic *meter* of water and add five *billion* tagged molecules: That's the same concentration, so we again expect

[7] For example, you'll use this skill to simulate bacterial genetics later in this chapter, and cellular mRNA populations in Chapter 8.
[8] See Section 3.2.2 (page 40).

$\langle \ell \rangle = 5$ for a sample of the same volume $V = 1\,\text{mm}^3$. Moreover, it seems reasonable that the entire distribution $\mathcal{P}(\ell)$ is essentially the same in this case as it was before. After all, each liter of that big thousand-liter bathtub has about five million tagged molecules, just as in the original situation. And in a $100\,\text{m}^3$ swimming pool, with $5 \cdot 10^{11}$ tagged molecules, the situation should be essentially the same. In short, it's reasonable to expect that there should be some *limiting distribution*, and that any large enough reservoir with concentration $c = 5 \cdot 10^6$ molecules per liter will give the same result for that distribution as any other. But "reasonable" is not enough. We need a proof. And anyway, we'd like to find an explicit formula for that limiting distribution.

4.3.2 The sum of many Bernoulli trials, each with low probability, follows a Poisson distribution

Translating the words of Section 4.3.1 into math, we are given values for the concentration c of tagged molecules and the sample volume V. We wish to find the distribution of the number ℓ of tagged molecules found in a sample, in the limit where the reservoir is huge but $\langle \ell \rangle$ is kept fixed. The discussion will involve several named quantities, so we summarize them here for reference:

V	sample volume, held fixed
V_*	reservoir volume, $\to \infty$ in the limit
$\xi = V/V_*$	probability that any one molecule is captured, $\to 0$ in the limit
c	concentration (number density), held fixed
$M_* = cV_*$	total number of molecules in the reservoir, $\to \infty$ in the limit
$\mu = cV = M_*\xi$	a constant as we take the limit
ℓ	number of tagged molecules in a particular sample, a random variable

Suppose that M_* molecules each wander through a reservoir of volume V_*, so $c = M_*/V_*$. We are considering a series of experiments all with the same concentration, so any chosen value of V_* also implies the value $M_* = cV_*$. Each molecule wanders independently of the others, so each has probability $\xi = V/V_* = Vc/M_*$ to be caught in the sample.

The total number caught thus reflects the sum of M_* identical, independent Bernoulli trials, whose distribution we have already worked out. Thus, we wish to compute

$$\lim_{M_* \to \infty} \mathcal{P}_{\text{binom}}(\ell; \xi, M_*), \text{ where } \xi = Vc/M_*. \tag{4.3}$$

The parameters V, c, and ℓ are to be held fixed when taking the limit.

Your Turn 4C

Think about how this limit implements the physical situation discussed in Section 4.3.1.

Notice that V and c enter our problem only via their product, so we will have one fewer symbol in our formulas if we eliminate them by introducing a new abbreviation $\mu = Vc$. The parameter μ is dimensionless, because the concentration c has dimensions of inverse volume (for example, "molecules per liter").

Substituting the Binomial distribution (Equation 4.1) into the expression above and rearranging gives

$$\lim_{M_* \to \infty} \left(\frac{\mu^\ell}{\ell!}\right) \left(1 - \frac{\mu}{M_*}\right)^{M_*} \left(1 - \frac{\mu}{M_*}\right)^{-\ell} \frac{M_*(M_* - 1) \cdots \big(M_* - (\ell - 1)\big)}{{M_*}^\ell}. \tag{4.4}$$

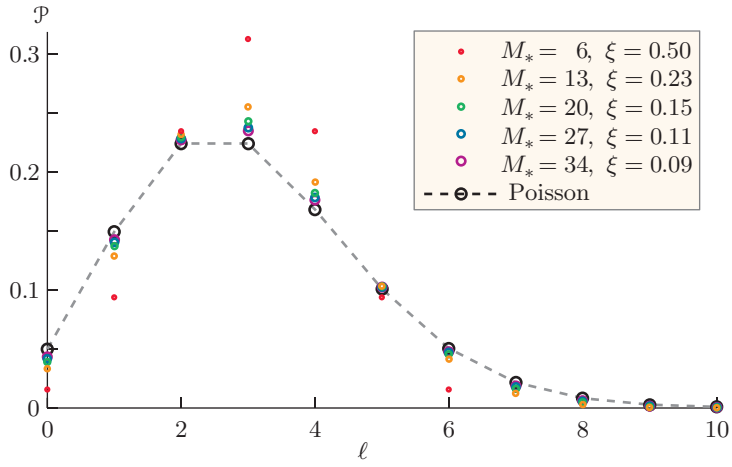


Figure 4.3 [Mathematical functions.] **Poisson distribution as a limit.** *Black circles* show the Poisson distribution for $\mu = 3$. The *dashed line* just joins successive points; the distribution is defined only at integer values of ℓ. The *colored circles* show how the Binomial distribution (Equation 4.3) converges to the Poisson distribution for large M_*, holding fixed $M_*\xi = 3$.

The first factor of expression 4.4 doesn't depend on M_*, so it may be taken outside of the limit. The third factor just equals 1 in the large-M_* limit, and the last one is

$$(1 - M_*^{-1})(1 - 2M_*^{-1}) \cdots \big(1 - (\ell - 1)M_*^{-1}\big).$$

Each of the factors above is very nearly equal to 1, and there are only $\ell - 1 \ll M_*$ of them, so in the limit the whole thing becomes another factor of 1, and may be dropped.

The second factor in parentheses in expression 4.4 is a bit more tricky, because its exponent is becoming large in the limit. To evaluate it, we need the compound interest formula:[9]

$$\lim_{M_* \to \infty} \left(1 - \frac{\mu}{M_*}\right)^{M_*} = \exp(-\mu). \tag{4.5}$$

To convince yourself of Equation 4.5, let $X = M_*/\mu$; then we want $\big((1 - X^{-1})^X\big)^\mu$. You can just evaluate the quantity $(1 - X^{-1})^X$ for large X on a calculator, and see that it approaches $\exp(-1)$. So the left side of Equation 4.5 equals e^{-1} raised to the power μ, as claimed.

Putting everything together then gives

$$\mathcal{P}_{\text{pois}}(\ell; \mu) = \frac{1}{\ell!}\, \mu^\ell e^{-\mu}. \quad \textbf{Poisson distribution} \tag{4.6}$$

Figure 4.3 illustrates the limit we have found, in the case $\mu = 3$. Figure 4.4 compares two Poisson distributions that have different values of μ. These distributions are not symmetric; for example, ℓ cannot be smaller than zero, but it can be arbitrarily large (because we took

[9] See page 20.

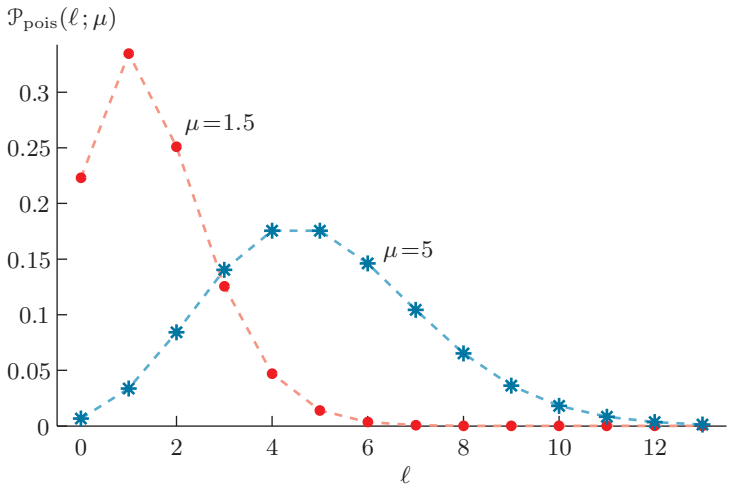


Figure 4.4 [Mathematical functions.] **Two examples of Poisson distributions.** Again, *dashed lines* just join successive points; Poisson distributions are defined only at integer values of ℓ.

the limit of large M_*). If μ is small, the distribution has a graph that is tall and narrow. For larger values of μ, the bump in the graph moves outward, and the distribution gets broader too.[10]

Your Turn 4D

Also graph the cases with $\mu = 0.1,\ 0.2$, and 1.

Example Confirm that the Poisson distribution is properly normalized for any fixed value of μ. Find its expectation and variance, as functions of the parameter μ.

Solution When we sum all the infinitely many entries in $\mathcal{P}_{\text{pois}}(\ell;\mu)$, we obtain $e^{-\mu}$ times the Taylor expansion for e^{μ} (see page 19). The product thus equals 1.

There are various ways to compute expectation and variance, but here is a method that will be useful in other contexts as well.[11] To find the expectation, we must evaluate $\sum_{\ell=0}^{\infty} \ell\mu^{\ell}e^{-\mu}/(\ell!)$. The trick is to start with the related expression $\frac{\mathrm{d}}{\mathrm{d}\mu}\left(\sum_{\ell=0}^{\infty}\mu^{\ell}/(\ell!)\right)$, evaluate it in two different ways, and compare the results.

On one hand, the quantity in parentheses equals e^{μ}, so its derivative is also e^{μ}. On the other hand, differentiating each term of the sum gives

$$\sum_{\ell=1}^{\infty} \ell\mu^{\ell-1}/(\ell!).$$

The derivative has pulled down a factor of ℓ from the exponential, making the expression almost the same as the quantity that we need.

[10] See Your Turn 4E.
[11] See Problem 7.2 and Section 5.2.4 (page 102).

Setting these two expressions equal to each other, and manipulating a bit, yields

$$1 = \mu^{-1}\left(\sum_{\ell=1}^{\infty} e^{-\mu}\ell\mu^{\ell}/(\ell!)\right) = \mu^{-1}\langle\ell\rangle.$$

Thus, $\langle\ell\rangle = \mu$ for the Poisson distribution with parameter μ. You can now invent a similar derivation and use it to compute var ℓ as a function of μ. [*Hint:* This time try taking *two* derivatives, in order to pull down two factors of ℓ from the exponent.]

Your Turn 4E

There is a much quicker route to the same answer. You have already worked out the expectation and variance of the Binomial distribution (the Example on page 72), so you can easily find them for the Poisson, by taking the appropriate limit (Equation 4.3). Do that, and compare your answer with the result computed directly in the Example just given.

To summarize,

- The Poisson distribution is useful whenever we are interested in *the sum of a lot of Bernoulli trials, each of which is individually of low probability.*
- In this limit, the two-parameter family of Binomial distributions collapses to a one-parameter family, a useful simplification in many cases where we know that M_* is large, but don't know its specific value.
- The expectation and variance have the key relationship

$$\text{var}\,\ell = \langle\ell\rangle \quad \text{for any Poisson distribution.} \tag{4.7}$$

4.3.3 Computer simulation

The method in Your Turn 4B can be used to simulate a Poisson-distributed random variable.[12] Although we cannot partition the unit interval into infinitely many bins, nevertheless in practice the Poisson distribution is very small for large ℓ, and so only a finite number of bins actually need to be set up.

4.3.4 Determination of single ion-channel conductance

Again, *the Poisson distribution is nothing new.* We got it as an approximation, a particular limiting case of the Binomial distribution. It's far more broadly applicable than it may seem from the motivating story in Section 4.3.1:

Whenever a large number of independent yes/no events each have low probability, but there are enough of them to ensure that the total "yes" count is nonnegligible, then that total will follow a Poisson distribution. (4.8)

[12] See Problem 4.6.

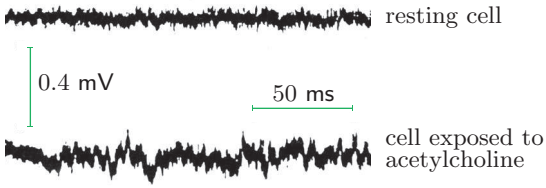


Figure 4.5 [Experimental data.] **Membrane electric potential in frog sartorius muscle.** The traces have been shifted vertically by arbitrary amounts; what the figure shows is the amplitude of the noise (randomness) in each signal. [From Katz & Miledi, 1972. ©Reproduced with permission of John Wiley & Sons, Inc.]

In the first half of the 20th century, it slowly became clear that cell membranes somehow could control their electrical conductance, and that this control lay at the heart of the ability of nerve and muscle cells to transmit information. One hypothesis for the mechanism of control was that the cell membrane is impermeable to the passage of ions (it is an insulator) but it is studded with tiny, discrete gateways. Each such gateway (or **ion channel**) can be open, allowing a particular class of ions to pass, or it can be shut. This switching, in turn, affects the electric potential across the membrane: Separating charges creates a potential difference, so allowing positive ions to reunite with negative ions reduces membrane potential.

The ion channel hypothesis was hotly debated, in part because at first, no component of the cell membrane was known that could play this role. The hypothesis made a prediction of the general magnitude of single-channel currents, but the prediction could not be tested: The electronic instrumentation of the day was not sensitive enough to detect the tiny postulated discrete electrical events.

B. Katz and R. Miledi broke this impasse, inferring the conductance of a single ion channel from a statistical analysis of the conductance of many such channels. They studied muscle cells, whose membrane conductance was known to be sensitive to the concentration of the neurotransmitter acetylcholine. Figure 4.5 shows two time series of the electric potential drop across the membrane of a muscle cell. The top trace is from a resting cell; the lower trace is from a muscle cell exposed to acetylcholine from a micropipette. Katz and Miledi noticed that the acetylcholine not only changed the resting potential but also increased the *noise* seen in the potential.[13] They interpreted this phenomenon by suggesting that the extra noise reflects independent openings and closings of a collection of many ion channels, as neurotransmitter molecules bind to and unbind from them.

In Problem 4.13, you'll follow Katz and Miledi's logic and estimate the effect of a single channel opening, from data similar to those in Figure 4.5. The experimenters converted this result into an inferred value of the channel conductance, which agreed roughly with the value expected for a nanometer-scale gateway, strengthening the ion channel hypothesis.

4.3.5 The Poisson distribution behaves simply under convolution

We have seen that the Poisson distribution has a simple relation between its expectation and variance. Now we'll find another nice property of this family of distributions, which also illustrates a new operation called "convolution."

[13] Other means of changing the resting potential, such as direct electrical stimulation, did not change the noisiness of the signal.

Example Suppose that a random variable ℓ is Poisson distributed with expectation μ_1, and m is another random variable, independent of ℓ, also Poisson distributed, but with a different expectation value μ_2. Find the probability distribution for the sum $\ell + m$, and explain how you got your answer.

Solution First, here is an intuitive argument based on physical reasoning: Suppose that we have blue ink molecules at concentration c_1 and red ink molecules at concentration c_2. A large chamber, of volume V_*, will therefore contain a total of $(c_1 + c_2)V_*$ molecules of either color. The logic of Section 4.3.2 then implies that the combined distribution is Poisson with $\mu = \mu_1 + \mu_2$.

Alternatively, here is a symbolic proof: First use the product rule for the independent variables ℓ and m to get the joint distribution $\mathcal{P}(\ell, m) = \mathcal{P}_{\text{pois}}(\ell; \mu_1)\mathcal{P}_{\text{pois}}(m; \mu_2)$. Next let $n = \ell + m$, and use the addition rule to find the probability that n has a particular value (regardless of the value of ℓ):

$$\mathcal{P}_{\text{n}}(n) = \sum_{\ell=0}^{n} \mathcal{P}_{\text{pois}}(\ell; \mu_1)\mathcal{P}_{\text{pois}}(n - \ell; \mu_2). \tag{4.9}$$

Then use the binomial theorem to recognize that this sum involves $(\mu_1 + \mu_2)^n$. The other factors also combine to give $\mathcal{P}_{\text{n}}(n) = \mathcal{P}_{\text{pois}}(n; \mu_1 + \mu_2)$.

Your Turn 4F

Again let $n = \ell + m$.
a. Use facts that you know about the expectation and variance of the Poisson distribution, and about the expectation and variance of a sum of independent random variables, to compute $\langle n \rangle$ and var n in terms of μ_1 and μ_2.
b. Now use the result in the Example above to compute the same two quantities and compare them with what you found in (a).

The right side of Equation 4.9 has a structure that arises in many situations,[14] so we give it a name: If f and g are any two functions of an integer, their **convolution** $f \star g$ is a new function, whose value at a particular n is

$$(f \star g)(n) = \sum_{\ell} f(\ell)g(n - \ell). \tag{4.10}$$

In this expression, the sum runs over all values of ℓ for which $f(\ell)$ and $g(n - \ell)$ are both nonzero. Applying the reasoning of the Example above to arbitrary distributions shows the significance of the convolution:

The distribution for the sum of two independent random variables is the convolution of their respective distributions. (4.11)

For the special case of Poisson distributions, the Example also showed that

The Poisson distributions have the special feature that the convolution of any two is again a Poisson distribution. (4.12)

[14] For example, see Sections 5.3.2 (page 108) and 7.5 (page 165). Convolutions also arise in image processing.

Your Turn 4G

Go back to Your Turn 3E (page 48). Represent the 36 outcomes of rolling two (distinct) dice as a 6×6 array, and circle all the outcomes for which the sum of the dice equals a particular value (for example, 6). Now reinterpret this construction as a convolution problem.

4.4 The Jackpot Distribution and Bacterial Genetics

4.4.1 It matters

Some scientific theories are pretty abstract. The quest to verify or falsify such theories may seem like a game, and indeed many scientists describe their work in those terms. But in other cases, it's clear right from the start that it matters a lot if a theory is right.

There was still active debate about the nature of inheritance at the turn of the 20th century, with a variety of opinions that we now caricature with two extremes. One pole, now associated with Charles Darwin, held that heritable changes in an organism arise spontaneously, and that evolution in the face of new environmental challenges is the result of selection applied to such mutation. The other extreme, now associated with J.-B. Lamarck, held that organisms actively create heritable changes in response to environmental challenges. The practical stakes could not have been higher. Josef Stalin imposed an agricultural policy based on the latter view that resulted in millions of deaths by starvation, and the near-criminalization of Darwinian theory in his country. The mechanism of inheritance is also critically important at the level of microorganisms, because the emergence of drug-resistant bacteria is a serious health threat today.

S. Luria and M. Delbrück set out to explore inheritance in bacteria in 1943. Besides addressing a basic biological problem, this work developed a key mode of scientific thought. The authors laid out two competing hypotheses, and sought to generate testable quantitative predictions from them. But unusually for the time, the predictions were *probabilistic* in character. No conclusion can be drawn from any single bacterium—sometimes it gains resistance; usually it doesn't. But the pattern of *large numbers* of bacteria has bearing on the mechanism. We will see how randomness, often dismissed as an unwelcome inadequacy of an experiment, turned out to be the most interesting feature of the data.

4.4.2 Unreproducible experimental data may nevertheless contain an important message

Bacteria can be killed by exposure to a chemical (for example, an antibiotic) or to a class of viruses called **bacteriophage** (abbreviated "phage"). In each case, however, some bacteria from a colony typically survive and transmit their resistance to their descendants. Even a colony founded from a *single* nonresistant individual will be found to have some resistant survivors. How is this possible?

Luria and Delbrück were aware that previous researchers had proposed both "Darwinian" and "Lamarckian" explanations for the acquisition of resistance, but that no fully convincing answer had been reached. They began their investigation by making the two alternatives more precise, and then drew predictions from them and

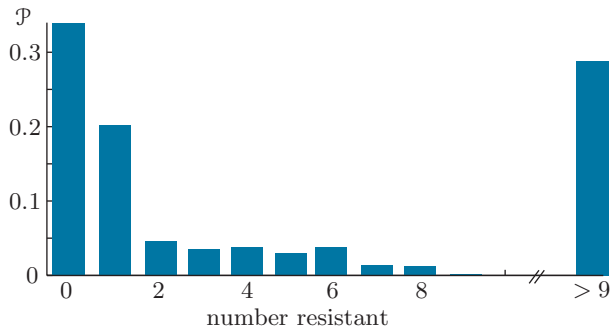


Figure 4.6 [Experimental data.] **Data from Luria and Delbrück's historic article.** This histogram represents one of their trials, consisting of 87 cultures. Figure 4.8 gives a more detailed representation of their experimental data and a fit to their model. [Data from Luria & Delbrück, 1943.]

designed an experiment intended to test the predictions. The Lamarckian hypothesis amounted to

> ***H1:*** *A colony descended from a single ancestor consists of identical individuals until a challenge to the population arises. When faced with the challenge, each individual struggles with it independently of the others, and most die. However, a small, randomly chosen subset of bacteria succeed in finding the change needed to survive the challenge, and are permanently modified in a way that they can transmit to their offspring.*

The Darwinian hypothesis amounted to

> ***H2:*** *No mutation occurs in* ***response*** *to the challenge. Instead, the entire colony is always spontaneously mutating, whether or not a challenge is presented. Once a mutation occurs, it is heritable. The challenge wipes out the majority, leaving behind only those individuals that had previously mutated to acquire resistance, and their descendants.*

In 1943, prior to the discovery of DNA's role in heredity, there was little convincing molecular basis for *either* of these hypotheses. An empirical test was needed.

Luria and Delbrück created a large collection of separate cultures of a particular strain of *Escherichia coli*. Each culture was given ample nutrients and allowed to grow for a time t_f, then challenged with a virus now called phage T1. To count the survivors, Luria and Delbrück spread each culture on a plate and continued to let them grow. Each surviving individual founded a colony, which eventually grew to a visible size. The survivors were few enough in number that these colonies were well separated, and so could be counted visually. Each culture had a different number m of survivors, so the experimenters reported not a single number but rather a histogram of the frequencies with which each particular value of m was observed (Figure 4.6).

Luria at once realized that the results were qualitatively unlike anything he had been trained to consider good science. In some ways, his data looked reasonable—the distribution had a peak near $m = 0$, then fell rapidly for increasing m. But there were also **outliers**, unexpected data points far from the main group.[15] Worse, when he performed the same

[15] Had Luria been content with two or three cultures, he might have missed the low-probability outliers altogether.

experiment a second and third time, the outliers, while always present, were quite different each time. It was tempting to conclude that this was just a bad, unreproducible experiment! In that case, the appropriate next step would be to work hard to find what was messing up the results (contamination?), or perhaps abandon the whole thing. Instead, Luria and Delbrück realized that hypothesis ***H2*** could explain their odd results.

The distributions we have encountered so far have either been exactly zero outside some range (like the Uniform and Binomial distributions), or at least have fallen off very rapidly outside a finite range (like Poisson or Geometric). In contrast, the empirical distribution in the Luria-Delbrück experiment is said to have a **long tail**; that is, the range of values at which it's nonnegligible extends out to very large m.[16] The more colorful phrase "jackpot distribution" is also used, by analogy to a gambling machine that generally gives a small payoff (or none), but occasionally gives a large one.

T_2 *Section 4.4.2′ (page 89) mentions one of the many additional tests that Luria and Delbrück made.*

4.4.3 Two models for the emergence of resistance

Luria and Delbrück reasoned as follows. At the start of each trial ("time zero"), a few nonresistant individuals are introduced into each culture. At the final time t_f, the population has grown to some large number $n(t_f)$; then it is subjected to a challenge, for example an attack by phage.

- ***H1*** states that each individual either mutates, with low probability ξ, or does not, with high probability $1 - \xi$, and that this random event is independent of every other individual. We have seen that in this situation, the total number m of individuals that succeed is distributed as a Poisson random variable. The data in Figure 4.6 don't seem to be distributed in this way.
- ***H2*** states that every time an individual divides, during the entire period from time zero to t_f, there is a small probability that it will spontaneously acquire the heritable mutation that confers resistance. So although the mutation event is once again a Bernoulli trial, according to ***H2*** it matters *when* that mutation occurred: Early mutants generate many resistant progeny, whereas mutants arising close to t_f don't have a chance to do so. Thus, in this situation there is an amplification of randomness.

Qualitatively, ***H2*** seems able to explain the observed jackpot distribution as a result of the occasional trial where the lucky mutant appeared early in the experiment (see Figure 4.7). A quantitative test is also required, however.

Note that both hypotheses contain a *single* unknown fitting parameter: in each case, a mutation probability. Thus, if we can adjust this one parameter to get a good fit under one hypothesis, but *no* value gives a good fit under the other hypothesis, then we will have made a fair comparison supporting the former over the latter. Note, too, that neither hypothesis requires us to understand the biochemical details of mutation, resistance, or inheritance. Both distill all of that detail into a single number, which is to be determined from data. If the winning model then makes *more* than one successful quantitative prediction (for example, if it predicts the entire shape of the distribution), then we may say that the data support it in a nontrivial way—they overconstrain the model.

[16] Some authors use the phrase "fat tail" to mean the same thing, because the tail of the distribution is larger numerically than we might have expected—it's "fat." Chapter 5 will give more examples illustrating the ubiquity of such distributions in Nature.

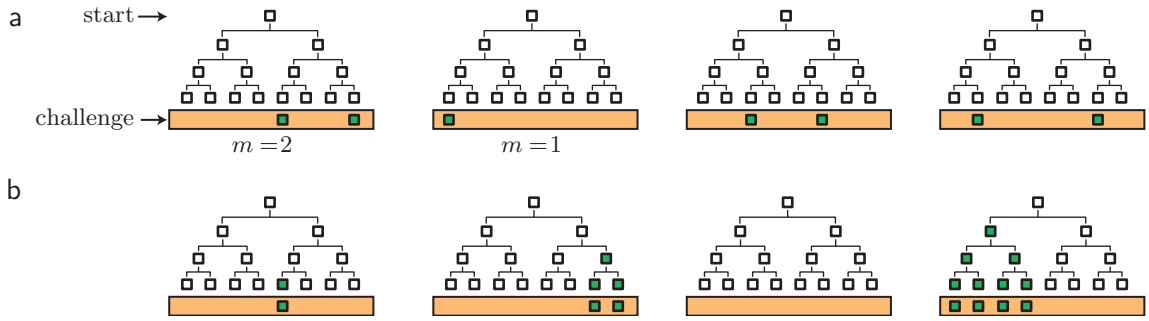


Figure 4.7 [Schematics.] **Two sets of imagined bacterial lineages relevant to the Luria-Delbrück experiment.** (a) The "Lamarckian" hypothesis states that bacterial resistance is created at the time of the challenge (*orange*). The number of resistant individuals (*green*) is then Poisson distributed. (b) The "Darwinian" hypothesis states that bacterial resistance can arise at any time. If it arises early (*second diagram*), the result can be very many resistant individuals.

T_2 *Section 4.4.3′ (page 89) gives more details about Luria and Delbrück's experiment.*

4.4.4 The Luria-Delbrück hypothesis makes testable predictions for the distribution of survivor counts

Hypothesis ***H1*** predicts that the probability distribution of the number of resistant bacteria is of the form $\mathcal{P}_{\text{poiss}}(m; \mu)$, where μ is an unknown constant. We need to find an equally specific prediction from ***H2***, in order to compare the two hypotheses. The discussion will involve several named quantities, so we summarize them here for reference:

n	cell population
g	number of doubling times (generations)
α_g	mutation probability per individual per doubling
t_f	final time
m	number of resistant mutant bacteria at time t_f
μ_{step}	expectation of number of new mutants in one doubling step
ℓ	number of new mutants actually arising in a particular doubling step, in a particular culture

Growth

Each culture starts at time zero with a known initial population n_0. (It's straightforward to estimate this quantity by sampling the bacterial suspension used to inoculate the cultures.) The growth of bacteria with plenty of food and no viral challenge can also be measured; it is exponential, doubling about every 25 minutes. Luria and Delbrück estimated $n_0 \approx 150$, and the final population to be $n(t_f) \approx 2.4 \cdot 10^8$. Thus, their growth phase consisted of $\log_2(2.4 \cdot 10^8/150) \approx 21$ doublings, a number we'll call g. We'll make the simplifying assumption that all individuals divide in synchrony, g times.

Mutation

Hypothesis ***H2*** states that, on every division, every individual makes an independent "decision" whether to make a daughter cell with the resistance mutation. Thus, the number of resistant individuals *newly arising on that division* is a Poisson-distributed random

variable whose expectation is proportional to the total population prior to that division. The constant of proportionality is the mutation probability per cell per doubling step, α_g, which is the one free parameter of the model. After mutation, the mutant cells continue to divide; we will assume that their doubling time is the same as that for the original-type cells.[17]

Computer simulation

In principle, we have now given enough information to allow a calculation of the expected Luria-Delbrück distribution $\mathcal{P}_{LD}(m; \alpha_g, n_0, g)$. In practice, however, it's difficult to do this calculation exactly; the answer is not one of the well-known, standard distributions. Luria and Delbrück had to resort to making a rather ad hoc mathematical simplification in order to obtain the prediction shown in Figure 4.6, and even then, the analysis was very involved. However, *simulating* the physical model described above with a computer is rather easy. Every time we run the computer code, we get a history of one simulated culture, and in particular a value for the final number m of resistant individuals. Running the code many times lets us build up a histogram of the resulting m values, which we can use either for direct comparison with experiment or for a calculation of reduced statistics like $\langle m \rangle$ or var m.

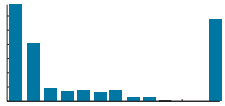


Figure 4.6 (page 82)

Such a simulation could work as follows. We maintain two population variables `Nwild` and `Nmutant`, with initial values n_0 and 0, respectively, and update them g times as follows. With each step, each population doubles. In addition, we draw a random number ℓ, representing the number of new mutants in that step, from a Poisson distribution with expectation $\mu_{step} = (\texttt{Nwild})\alpha_g$, then add ℓ to `Nmutant` and subtract it from `Nwild`. The final value of `Nmutant` after g doubling steps gives m for that simulated culture. We repeat many times for one value of the parameter α_g, compare the resulting probability distribution with experimental data, then adjust α_g and try again until we are satisfied with the fit (or convinced that no value of α_g is satisfactory).

The strategy just outlined points out a payoff for our hard work in Section 4.3. One could imagine simply simulating `Nwild` Bernoulli trials in each doubling step. But with hundreds of millions of individuals to be polled in the later steps, we'd run out of computing resources! Because all we really need is the *number* of mutants, we can instead make a *single* draw from a Poisson distribution for each doubling step.

Results

Problem 4.14 gives more details on how to carry out these steps. Figure 4.8a shows data from the experiment, together with best-fit distributions for each of the two hypotheses. It may not be immediately apparent from this presentation just how badly ***H1*** fails. One way to see the failure is to note that the experimental data have sample mean $\bar{m} \approx 30$ but variance ≈ 6000, inconsistent with any Poisson distribution (and hence with hypothesis ***H1***).[18]

The figure also shows that ***H2*** does give a reasonable account of the entire distribution, with only one free fit parameter, whereas ***H1*** is unable to explain the existence of *any* cultures having more than about five mutants. To bring this out, Figure 4.8b shows the same information as panel (a) but on a logarithmic scale. This version also shows that the deviation of ***H1*** at $m = 20$ from the experimental observation is far more significant than that of ***H2*** at $m = 4$.

---

[17] T_2 See Section 4.4.3′ (page 89).

[18] See Problem 4.15.

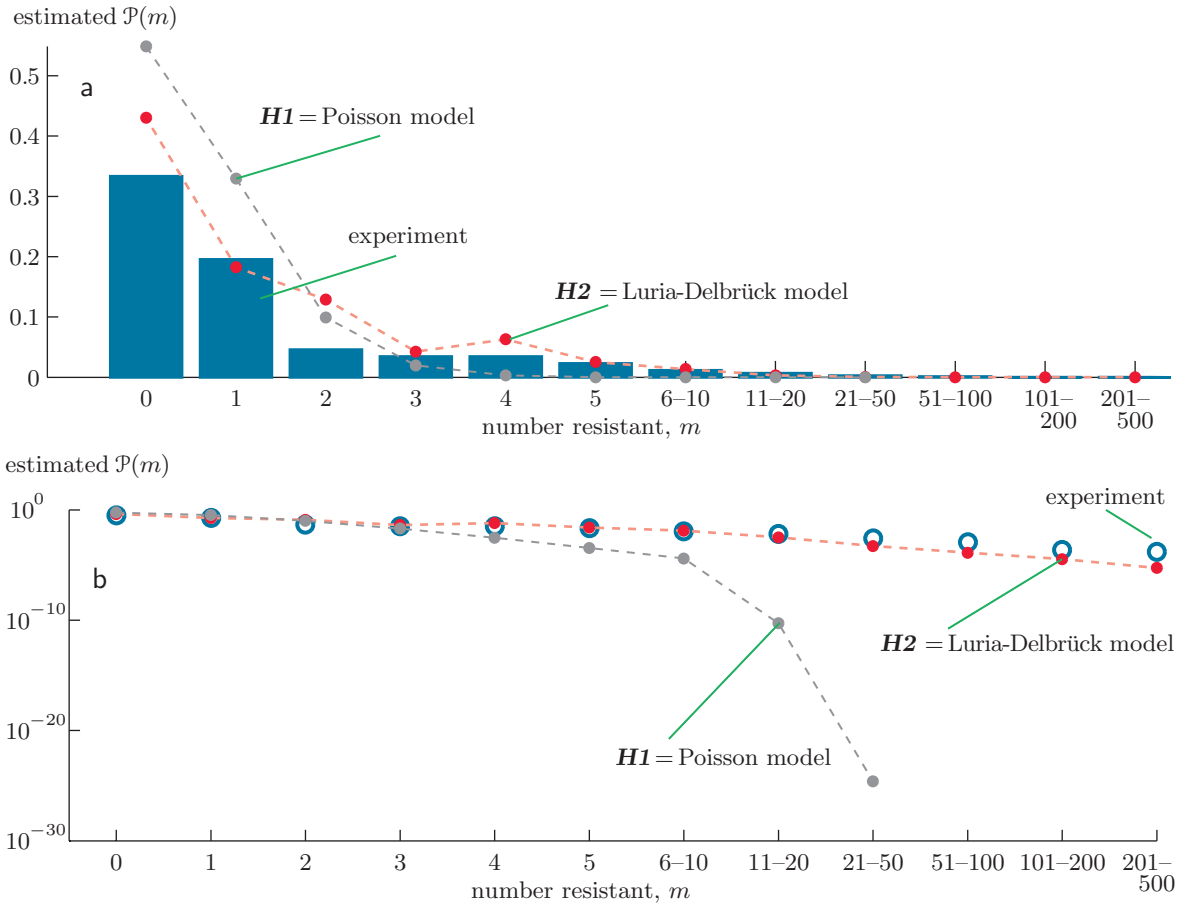


Figure 4.8 [Experimental data with fits.] **Two models compared to data on acquired resistance.** (a) *Bars:* Data from the same experiment as in Figure 4.6. The *gray dots* show a fit to data under the "Lamarckian" hypothesis ***H1***. The *red dots* show a fit under the Luria-Delbrück hypothesis ***H2***. (b) The same as (a), plotted in semilog form to highlight the inability of ***H1*** to account for the outliers in the data. Luria and Delbrück combined the data for high mutant number m by lumping several values together, as indicated in the horizontal axis labels. Both panels correct for this: When a bin contains K different values of m all lumped together, its count has been divided by K, so that the bar heights approximate the probabilities for individual values of m. That is, each bar represents the estimated $\mathcal{P}(m)$ for single values of m.

Your Turn 4H

Figure 4.6 appears to show two bumps in the probability, whereas Figure 4.8a does not. Explain this apparent discrepancy.

4.4.5 Perspective

Luria and Delbrück's experiment and analysis showed dramatically that bacteriophage resistance was the result of spontaneous mutation, not the survival challenge itself. Similar mechanisms underlie other evolutionary phenomena, including viral evolution in a single HIV patient, discussed in the Prolog to this book.[19]

[19] Also see Problem 3.6.

This work also provided a framework for the quantitative measurement of extremely low mutation probabilities. Clearly α_g must be on the order of 10^{-8}, because hundreds of millions of bacteria contain only a handful of resistant mutants. It may be mind-boggling to imagine checking all the population and somehow counting the resistant members, but the authors' clever experimental method accomplished just that. At first, this counting seemed to give contradictory results, due to the large spread in the result m. Then, however, Luria and Delbrück had the insight of making a *probabilistic* prediction, comparing it to *many* trials, and finding the *distribution* of outcomes. Fitting that distribution did lead to a good measurement of α_g. As Luria and Delbrück wrote, "The quantitative study of bacterial variation has [until now] been hampered by the apparent lack of reproducibility of results, which, as we show, lies in the very nature of the problem and is an essential element for its analysis."

Your dull-witted but extremely fast assistant was a big help in this analysis. Not every problem has such a satisfactory numerical solution, just as not every problem has an elegant analytic (pencil-and-paper) solution. But the set of problems that are easy analytically, and that of problems that are easy numerically, are two different domains. Scientists with both kinds of toolkit can solve a broader range of problems.

T_2 *Section 4.4.5′ (page 89) discusses some qualifications to the Darwinian hypothesis discussed in this chapter, in the light of more recent discoveries in bacterial genetics, as well as an experiment that further confirmed Luria and Delbrück's interpretation.*

THE BIG PICTURE

Many physical systems generate partially random behavior. If we treat the distribution of outcomes as completely unknown, then we may find it unmanageable, and uninformative, to determine that distribution empirically. In many cases, however, we can formulate some well-grounded expectations that narrow the field considerably. From such "insider information"—a model—we can sometimes predict most of the behavior of a system, leaving only one or a few parameter values unknown. Doing so not only lightens our mathematical burden; it can also make our predictions specific, to the point where we may be able to falsify a hypothesis by embodying it in a model, and showing that *no* assumed values of the parameters make successful predictions.

For example, it was reasonable to suppose that a culture of bacteria suspended in liquid will all respond independently of each other to attack by phage or antibiotic. From this assumption, Luria and Delbrück got falsifiable predictions from two hypotheses, and eliminated one of them.

Chapter 6 will start to systematize the procedure for simultaneously testing a model and determining the parameter values that best represent the available experimental data. First, however, we must extend our notions of probability to include continuously varying quantities (Chapter 5).

KEY FORMULAS

- *Binomial distribution:* $\mathcal{P}_{\text{binom}}(\ell;\xi,M) = \frac{M!}{\ell!(M-\ell)!}\,\xi^{\ell}(1-\xi)^{M-\ell}$. The random variable ℓ is drawn from the sample space $\{0, 1, \ldots, M\}$. The parameters ξ and M, and $\mathcal{P}$ itself, are all dimensionless. The expectation is $\langle\ell\rangle = M\xi$, and the variance is $\text{var}\,\ell = M\xi(1-\xi)$.
- *Simulation:* To simulate a given discrete probability distribution $\mathcal{P}$ on a computer, divide the unit interval into bins of widths $\mathcal{P}(\ell)$ for each allowed value of ℓ. Then choose Uniform random numbers on that interval and assign each one to its appropriate bin. The resulting bin assignments are draws from the desired distribution.

- *Compound interest:* $\lim_{M\to\infty}\big(1 \pm (a/M)\big)^M = \exp(\pm a)$.
- *Poisson distribution:* $\mathcal{P}_{\text{pois}}(\ell;\mu) = \mathrm{e}^{-\mu}\mu^\ell/(\ell!)$. The random variable ℓ is drawn from the sample space $\{0, 1, \dots\}$. The parameter μ, and $\mathcal{P}$ itself, are both dimensionless. The expectation and variance are $\langle\ell\rangle = \operatorname{var}\ell = \mu$.
- *Convolution:* $(f\star g)(m) = \sum_\ell f(\ell)g(m-\ell)$. Then $\mathcal{P}_{\text{pois}}(\bullet;\mu_1)\star\mathcal{P}_{\text{pois}}(\bullet;\mu_2) = \mathcal{P}_{\text{pois}}(\bullet;\mu_{\text{tot}})$, where $\mu_{\text{tot}} = \mu_1 + \mu_2$.

FURTHER READING

Semipopular:
Discovery of phage viruses: Zimmer, 2011.
On Delbrück and Luria: Luria, 1984; Segrè, 2011. Long-tail distributions: Strogatz, 2012.

Intermediate:
Luria-Delbrück: Benedek & Villars, 2000, §3.5; Phillips et al., 2012, chapt. 21.

Technical:
Luria & Delbrück, 1943.
Estimate of ion channel conductance: Bialek, 2012, §2.3.
Calibration of fluorescence by Binomial partitioning: Rosenfeld et al., 2005, supporting online material.

T_2 **Track 2**

4.4.2′ On resistance

Our model of the Luria-Delbrück experiment assumed that the resistant cells were like the wild type, except for the single mutation that conferred resistance to phage infection. Before concluding this, Luria and Delbrück had to rule out an alternative possibility to be discussed in Chapter 10, that their supposedly resistant cells had been transformed to a "lysogenic" state. They wrote, "The resistant cells breed true . . . . No trace of virus could be found in any pure culture of the resistant bacteria. The resistant strains are therefore to be considered as non-lysogenic."

T_2 **Track 2**

4.4.3′ More about the Luria-Delbrück experiment

The discussion in the main text hinged on the assumption that initially the cultures of bacteria contained no resistant individuals. In fact, any colony could contain such individuals, but only at a very low level, because the resistance mutation also slows bacterial growth. Luria and Delbrück estimated that fewer than one individual in 10^5 were resistant. They concluded that inoculating a few dozen cultures, each with a few dozen individuals, was unlikely to yield even one culture with one resistant individual initially.

The analysis in Section 4.4.4 neglected the reproduction penalty for having the resistance mutation. However, the penalty needed to suppress the population of initially resistant individuals is small enough not to affect our results much. If we wish to do better, it is straightforward to introduce two reproduction rates into the simulation.

T_2 **Track 2**

4.4.5′a Analytical approaches to the Luria-Delbrück calculation

The main text emphasized the power of computer simulation to extract probabilistic predictions from models such as Luria and Delbrück's. However, analytic methods have also been developed for this model as well as for more realistic variants (Lea & Coulson, 1949; Rosche & Foster, 2000).

4.4.5′b Other genetic mechanisms

The main text outlined a turning point in our understanding of genetics. But our understanding continues to evolve; no one experiment settles everything forever. Thus, the main text didn't say "inheritance of acquired characteristics is wrong"; instead, we outlined how one specific implementation of that idea led to quantitatively testable predictions about one particular system, which were falsified.

Other mechanisms of heritable change have later been found that are different from the random mutations discussed in the main text. For example,

- A virus can integrate its genetic material into a bacterium and lie dormant for many generations ("lysogeny"; see Chapter 10).
- A virus can add a "plasmid," a small autonomous loop of DNA that immediately confers new abilities on its host bacterium without any classical mutation, and that is copied and passed on to offspring.

- Bacteria can also exchange genetic material among themselves, with or without help from viruses ("horizontal gene transfer"; see Thomas & Nielsen, 2005).
- Genetic mutations themselves may not be uniform, as assumed in neo-Darwinian models: Regulation of mutation rates can itself be an adaptive response to stress, and different loci on the genome have different mutation rates.

None of these mechanisms should be construed as a failure of Darwin's insight, however. Darwin's framework was quite general; he did not assume Mendelian genetics, and in fact was unaware of it. Instead, we may point out that the mechanisms listed above that lie outside of classical genetics reflect competencies that cells possess *by virtue of their genetic makeup,* which itself evolves under natural selection.

4.4.5′c Non-genetic mechanisms

An even broader class of heritable but non-genetic changes has been found, some of which are implicated in resistance to drug or virus attack:

- The available supply of nutrients can "switch" a bacterium into a new state, which persists into its progeny, even though no change has occurred to the genome (again see Chapter 10). Bacteria can also switch spontaneously, for example, creating a subpopulation of slowly growing "persistors" that are resistant to antibiotic attack. Such "epigenetic" mechanisms (for example, involving covalent modifications of DNA without change to its sequence) have also been documented in eukaryotes.
- Clustered regularly interspaced short palindromic repeats (CRISPR) have been found to give a nearly "Lamarckian" mechanism of resistance (Barrangou et al., 2007; Koonin & Wolf, 2009).
- Drug and virus resistance have also been documented via gene silencing by RNA interference (Calo et al., 2014 and Rechavi, Minevich, and Hobert 2011; see also Section 9.3.3′, page 234).

4.4.5′d Direct confirmation of the Luria-Delbrück hypothesis

The main text emphasized testing a hypothesis based on its probabilistic predictions, but eight years after Luria and Delbrück's work it became possible to give a more direct confirmation. J. Lederberg and E. Lederberg created a bacterial culture and spread it on a plate, as usual. Prior to challenging the bacteria with antibiotic, however, they let them grow on the plate a bit longer, then *replicated* the plate by pressing an absorbent fabric onto it and transferring it to a second plate. The fabric picked up some of the bacteria in the first plate, depositing them in the same relative positions on the second. When both plates were then subjected to viral attack, they showed colonies of resistant individuals in corresponding locations, demonstrating that those subcolonies existed prior to the attack (Lederberg & Lederberg, 1952).

PROBLEMS

4.1 Risk analysis

In 1941, the mortality (death) rate for 75-year-old people in a particular region of the United States was 0.089 per year. Ten thousand people of this age were all given a vaccine, and one died within 12 hours. Should this be attributed to the vaccine? Calculate the probability that at least one would have died in 12 hours, *even without* the vaccine.

4.2 Binning jitter

Here is a more detailed version of Problem 3.3. Nora asked a computer to generate 3000 Uniformly distributed, random binary fractions, each six bits long (see Equation 3.1, page 36), and made a histogram of the outcomes, obtaining Figure 4.9. It doesn't look very Uniform. Did Nora (or her computer) make a mistake? Let's investigate.

a. Qualitatively, why isn't it surprising that the bars are not all of equal height?
   Now get more quantitative. Consider the first bar, which represents the binary fraction corresponding to 000000. The probability of that outcome is 1/64. The computer made 3000 such draws and tallied how many had this outcome. Call that number N_{000000}.

b. Compute the expectation, variance, and standard deviation of N_{000000}.

c. The height of the first bar is $N_{000000}/3000$. Compute the standard deviation of this quantity. The other bars will also have the same standard deviation, so comment on whether your calculated value appears to explain the behavior seen in the figure.

4.3 Gene frequency

Consider a gene with two possible variants (alleles), called *A* and *a*.

Father Fish has two copies of this gene in every somatic (body) cell; suppose that each cell has one copy of allele *A* and one copy of *a*. Father fish makes a zillion sperm, each with just one copy of the gene. Mother Fish also has genotype *Aa*. She makes a zillion eggs, again each with just one copy of the gene.

Four sperm and four eggs are drawn at random from these two pools and fuse, giving four fertilized eggs, which grow as usual.

a. What is the total number of copies of *A* in these four fertilized eggs? Re-express your answer in terms of the "frequency of allele *A*" in the new generation, which is the total number of copies of *A* in these four individuals, divided by the total number of either *A* or *a*. Your answer should be a symbolic expression.

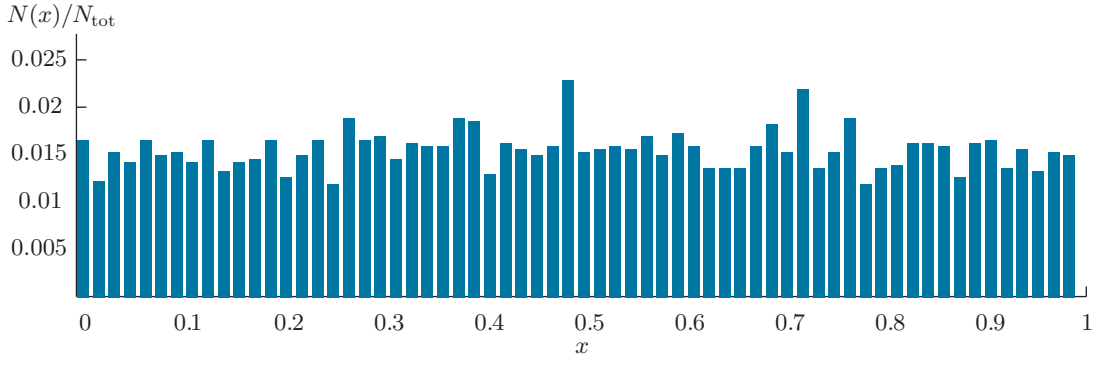


Figure 4.9 [Simulated data.] See Problem 4.2.

b. What is the probability that the frequency of allele A is exactly the same in the new generation as it was in the parent generation? Your answer should be a number. What is the probability that the frequency of allele A is *zero* in the new generation?

4.4 Partitioning error

Idealize a dividing bacterium as a well-mixed box of molecules that suddenly splits into two compartments of equal volume. Suppose that, prior to the division, there are 10 copies of a small, mobile molecule of interest to us. Will we always get exactly 5 on each side? If not, how probable is it that one side, or the other, will get 3 or fewer copies?

4.5 More about random walks

If you haven't done Problem 3.4 yet, do it before starting this problem. Again set up a simulation of a random walk, initially in a single dimension x with steps of length $d = 1\,\mu\text{m}$. Let x_* be the location relative to the origin at time $t = 20\,\text{s}$. It's a random variable, because it's different each time we make a new trajectory.

a. Compute $\langle (x_*)^2 \rangle$.

b. Let x_{**} be the location at $t = 40\,\text{s}$, and again compute $\langle (x_{**})^2 \rangle$.

c. Now consider a *two*-dimensional random walk: Here the chess piece makes moves in a *plane*. The x and y components of each move are each random, and independent of each other. Again, x steps by $\pm 1\ \mu\text{m}$ in each move; y has the same step distribution. Find $\langle (r_*)^2 \rangle$ for this situation, where $r^2 = x^2 + y^2$ and again the elapsed time is 20 s.

d. Returning to the one-dimensional walker in part (a), this time suppose that it steps in the + direction 51% of the time and in the − direction 49% of the time. What are the expectation and variance of x_* in this situation?

4.6 Simulate a Poisson distribution

a. Write a function for your computer called `poissonSetup(mu)`, similar to the one described in Your Turn 4B, but which prepares a set of bin edges suitable for simulating a Poisson distribution with expectation `mu`. In principle, this distribution has infinitely many bins, but in practice you can cut it off; that is, use either 10 or 10`mu` bins (rounded to an integer), whichever is larger. (Or you may invent a more clever way to find a suitable finite cutoff.)

b. Write a little "wrapper" program that calls `poissonSetup(2)`, and then generates 10 000 numbers from the distribution, finds the sample mean and variance, and histograms the distribution.

c. Repeat with `mu` = 20, and comment on the different symmetry of the peak between this case and (b). Confirm that the sample mean and variance you found agree with direct calculation from the definition of the Poisson distribution.

4.7 Simulate a Geometric distribution

Do Problem 4.6, but with Geometric instead of Poisson distributions. Try the cases with $\xi = 1/2$ and $1/20$.

4.8 Cultures and colonies

a. Suppose that you add $2 \cdot 10^8$ virions to a culture containing 10^8 cells. Suppose that every virus "chooses" a cell at random and successfully infects it, but some mechanism prevents infected cells from lysing. A cell can be infected by more than one virus, but suppose that a prior infection doesn't alter the probability of another one. What fraction of the cells

will remain uninfected? How many virions would have been required had you wished for over 99% of the cells in the culture to be infected?

b. Suppose that you take a bacterial culture and dilute it by a factor of one million. Then you spread 0.10 mL of this well-mixed, diluted culture on a nutrient plate, incubate, and find 110 well-separated colonies the next day. What was the concentration of live bacteria (colony forming units, or CFU) in the original culture? Express your answer as CFU/mL and also give the standard deviation of your estimate.

4.9 Poisson limit

The text argued analytically that the Poisson distribution becomes a "good" approximation to the Binomial distribution in a certain limiting case. Explore the validity of the argument:

a. Compute the natural log of the Binomial distribution with $M = 100$ and $\xi = 0.03$, at all values of ℓ. Compare the log of the corresponding Poisson distribution by graphing both. Make another graph showing the actual value (not the log) of each distribution for a range of ℓ values close to $M\xi$.

b. Repeat, but this time use $\xi = 0.5$.

c. Repeat, but this time use $\xi = 0.97$.

d. Comment on your results in the light of the derivation in Section 4.3.2 (page 75).

[*Hint:* Your computer math package may be unable to compute quantities like 100! directly. But it will have no difficulty computing $\ln(1) + \cdots + \ln(100)$. It may be particularly efficient to start with $\ell = 1$, where $\mathcal{P}_{\text{binom}}$ and $\mathcal{P}_{\text{pois}}$ are both simple, then obtain each succeeding $\mathcal{P}(\ell)$ value from its predecessor.]

4.10 Cancer clusters

Obtain Dataset 6. The variable `incidents` contains a list of (x, y) coordinate pairs, which we imagine to be the geographic coordinates of the homes of persons with some illness.

a. First create a graphical representation of these points. The variable `referencepoints` contains coordinates of four landmarks; add them to your plot in a different color.

Suppose that someone asks you to investigate the cause of that scary cluster near reference point #3, and the relative lack of cases in some other regions. Before you start looking for nuclear reactors or cell-phone towers, however, the first thing to check is the "null hypothesis": Maybe these are just points randomly drawn from a Uniform distribution. There's no way to prove that a single instance of dots "is random." But we can try to make a quantitative prediction from the hypothesis and then check whether the data roughly obey it.[20]

b. Add vertical lines to your plot dividing it into N equal strips, either with your computer or by drawing on a hard copy of your plot. Choose a value of N somewhere between 10 and 20. Also add the same number of horizontal lines dividing it into N equal strips. Thus, you have divided your graph into a grid of N^2 blocks. (What's wrong with setting up fewer than 100 blocks? What's wrong with more than 400?)

c. Count how many dots lie in each of the blocks. Tally up how many blocks have 0, 1, ... dots in them. That gives you the frequency $F(\ell)$ to find ℓ dots in a block, and hence an estimate for the probability $\mathcal{P}_{\text{est}}(\ell) = F(\ell)/N^2$ that a block will have ℓ dots. The dataset contains a total of 831 points, so the average number of dots per block is $\mu = 831/N^2$.

[20] The next step would be to obtain new data and see if the same hypothesis, *with no further tweaking*, also succeeds on them, but this is not always practical.

d. If we had a huge map with lots of blocks, and dots distributed Uniformly and independently over that map with an average of μ per block, then the actual number observed in a block would follow a known distribution. Graph this probability distribution for a relevant range of ℓ values. Overlay a graph of the estimated distribution $\mathcal{P}_{\text{est}}$ that you obtained in (c). Does the resulting picture seem to support the null hypothesis?

e. For comparison, generate 831 simulated data points that really are Uniformly distributed over the region shown, and repeat the above steps.

4.11 Demand fluctuations

In a large fleet of delivery trucks, the average number inoperative on any day, due to breakdowns, is two. Some standby trucks are also available. Find numerical answers for the probability that on any given day

a. No standby trucks are needed.

b. More than one standby truck is needed.

4.12 Low probability

a. Suppose that we have an unfair "coin," for which flipping heads is a rather rare event, a Bernoulli trial with $\xi = 0.08$. Imagine making $N = 1000$ trials, each consisting of 100 such coin flips. Write a computer simulation of such an experiment, and for each trial compute the total number of heads that appeared. Then plot a histogram of the frequencies of various outcomes.

b. Repeat for $N = 30\,000$ and comment. What was the most frequent outcome?

c. Superimpose on the plot of (a) the function $1000\mathcal{P}_{\text{pois}}(\ell; 8)$, and compare the two graphs.

4.13 Discreteness of ion channels

Section 4.3.4 introduced Katz and Miledi's indirect determination of the conductance of a single ion channel, long before biophysical instruments had developed enough to permit direct measurement. In this problem, you'll follow their logic with some simplifying assumptions to make the math easier.

For concreteness, suppose that each channel opening causes the membrane to depolarize slightly, increasing its potential by an amount a for a fixed duration τ; afterward the channel closes again. There are M channels; suppose that M is known to be very large. Each channel spends a small fraction ξ of its time open in the presence of acetylcholine, and all channels open and close independently of one another. Suppose also that when ℓ channels are simultaneously open, the effect is linear (the excess potential is ℓa).

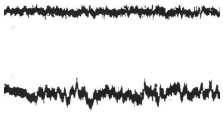


Figure 4.5 (page 79)

a. None of the parameters a, τ, M, or ξ is directly measurable from data like those in Figure 4.5. However, two quantities *are* measurable: the mean and the variance of the membrane potential. Explain why the Poisson distribution applies to this problem, and use it to compute these quantities in terms of the parameters of the system.

b. The top trace in the figure shows that even in the resting state, where all the channels are closed, there is still some electrical noise for reasons unrelated to the hypothesis being considered. Explain why it is legitimate to simply subtract the average and variance of this resting-state signal from that seen in the lower trace.

c. Show how Katz and Miledi's experimental measurement of the change in the average and the variance of the membrane potential upon acetylcholine application allows us to deduce the value of a. (This value is the desired quantity, a measure of the effect of a single channel opening; it can be converted to a value for the conductance of a single channel.)

d. In a typical case, Katz and Miledi found that the average membrane potential increased by 8.5 mV and that the variance increased by $(29.2\ \mu\text{V})^2$ after application of acetylcholine. What then was a?

4.14 Luria-Delbrück experiment

First do Problem 4.6, and be sure that your code is working the way you expect before attempting this problem.

Imagine a set of C cultures (separate flasks) each containing n_0 bacteria initially. Assume that all the cells in a culture divide at the same time, and that every time a cell divides, there is a probability α_g that one of the daughter cells will mutate to a form that is resistant to phage attack. Assume that the initial population has no resistant mutants ("pure wild-type"), and that all progeny of resistant cells are resistant ("no reversion"). Also assume that mutant and wild-type bacteria multiply at the same rate (no "fitness penalty"), and that at most one of the two daughter cells mutate (usually neither).

a. Write a computer code to simulate the situation and find the number of resistant mutant cells in a culture after g doublings. The Poisson distribution gives a good approximation to the number of new mutants after each doubling, so use the code you wrote in Problem 4.6. Each simulated culture will end up with a different number m of resistant mutant cells, due to the random character of mutation.

b. For $C = 500$ cultures with $n = 200$ cells initially, and $\alpha_g = 2 \cdot 10^{-9}$, find the number of cultures with m resistant mutant cells after $g = 21$ doublings, as a function of m. Plot your result as an estimated probability distribution. Compare its sample mean to its variance and comment.

c. Repeat the simulation $M = 3$ times (that is, M sets of C cultures), and comment on how accurately we can expect to find the true expectation and variance of the distribution from such experiments.

d. The chapter claimed that the origin of the long tail in the distribution is that on rare occasions a resistant mutant occurs earlier than usual, and hence has lots of offspring. For each simulated culture, let i_* denote at which step (number of doublings) the *first* mutant appears (or $g + 1$ if never). Produce a plot with m on one axis and i_* on the other, and comment.

[*Hints:* (*i*) This project will require a dozen or so lines of code, more complex than what you've done so far. Outline your algorithm before you start to code. Keep a list of all the variables you plan to define, and give them names that are meaningful to you. (You don't want two unrelated variables both named `n`.)
(*ii*) Start with smaller numbers, like $C = 100, M = 1$, so that your code runs fast while you're debugging it. When it looks good, then substitute the requested values of those parameters.
(*iii*) One way to proceed is to use three nested loops: The outermost loop repeats the code for each simulated experiment, from 1 to M. The middle loop involves which culture in a particular experiment is being simulated, from 1 to C. The innermost loop steps through the doublings of a particular experiment, in a particular culture.[21]
(*iv*) Remember that in each doubling step the only candidates for mutation are the remaining unmutated cells.]

4.15 Luria-Delbrück data

a. Obtain Dataset 5, which contains counts of resistant bacteria in two of the Luria-Delbrück experiments. For their experiment #23, find the sample mean and variance in the number

[21] More efficient algorithms are possible.

of resistant mutants, and comment on the significance of the values you obtain. [*Hint:* The count data are presented in bins of nonuniform size, so you'll need to correct for that. For example, five cultures were found to have between 6 and 10 mutants, so assume that the five instances were spread uniformly across those five values (in this case, one each with 6, 7, 8, 9, and 10 mutants).]

b. Repeat for their experiment #22.

4.16 T_2 Skewed distribution

Suppose that ℓ is drawn from a Poisson distribution. Find the expectation $\left\langle \left(\ell - \langle \ell \rangle\right)^3 \right\rangle$, which depends on μ. Compare your answer with the case of a symmetric distribution, and suggest an interpretation of this statistic.

Continuous Distributions

The generation of random numbers is too important to be left to chance.
—Robert R. Coveyou

5.1 Signpost

Some of the quantities that we measure are discrete, and the preceding chapters have used discrete distributions to develop many ideas about probability and its role in physics, chemistry, and biology. Most measured quantities, however, are inherently continuous, for example, lengths or times.[1] Figure 3.2b showed one attempt to represent the distribution of such a quantity (a waiting time) by artificially dividing its range into bins, but Nature does not specify any such binning. In other cases, a random quantity may indeed be discrete, but with a distribution that is roughly the same for neighboring values, as in Figure 3.1b; treating it as continuous may eliminate an irrelevant complication.

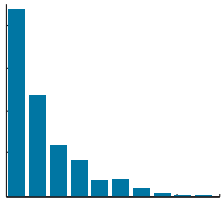


Figure 3.2b (page 38)

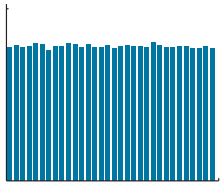


Figure 3.1b (page 37)

This chapter will extend our previous ideas to the continuous case. As in the discrete case, we will introduce just a few standard distributions that apply to many situations that arise when we make physical models of living systems.

This chapter's Focus Question is

Biological question: What do neural activity, protein interaction networks, and the diversity of antibodies all have in common?

Physical idea: Power-law distributions arise in many biophysical contexts.

[1] Some authors call a continuous random variable a "metric character," in distinction to the discrete case ("meristic characters").

5.2 Probability Density Function

5.2.1 The definition of a probability distribution must be modified for the case of a continuous random variable

In parallel with Chapter 3, consider a replicable random system whose samples are described by a continuous quantity x—a **continuous random variable**. x may have dimensions. To describe its distribution, we *temporarily* partition the range of allowed values for x into bins of width Δx, each labeled by the value of x at its center. As in the discrete case, we again make many measurements and find that ΔN of the N measurements fall in the bin centered on x_0, that is, in the range from $x_0 - \frac{1}{2}\Delta x$ to $x_0 + \frac{1}{2}\Delta x$. The integer ΔN is the frequency of the outcome.

We may be tempted now to define $\wp(x_0) \stackrel{?}{=} \lim_{N_{\text{tot}}\to\infty} \Delta N/N_{\text{tot}}$, as in the discrete case. The problem with this definition is that, in the limit of small Δx, it always goes to *zero*—a correct but uninformative answer. After all, the fraction of students in a class with heights between, say, 199.999 999 and 200.000 001 cm is very nearly zero, regardless of how large the class is. More generally, we'd like to invent a description of a continuous random system that doesn't depend on any extrinsic choice like a bin width.

The problem with the provisional definition just proposed is that when we cut the bin width in half, each of the resulting half-bins will contain roughly half as many observations as previously.[2] To resolve this problem, in the continuous case we modify the provisional definition of probability distribution by introducing a factor of $1/(\Delta x)$. Dividing by the bin width has the desirable effect that, if we subdivide each bin into two, then we get canceling factors of 1/2 in numerator and denominator, and no net change in the quotient. Thus, at least in principle, we can keep reducing Δx until we obtain a continuous function of x, at the value x_0:

$$\wp_{\text{x}}(x_0) = \lim_{\Delta x\to 0}\left(\lim_{N_{\text{tot}}\to\infty}\frac{\Delta N}{N_{\text{tot}}\Delta x}\right). \tag{5.1}$$

As with discrete distributions, we may drop the subscript "x" if the value of x completely describes our sample space, or more generally if this abbreviation will not cause confusion.

Even if we have only a finite number of observations, Equation 5.1 gives us a way to make an estimate of the pdf from data:

> *Given many observations of a continuous random variable x, choose a set of bins that are narrow, yet wide enough to each contain many observations. Find the frequencies ΔN_i for each bin centered on x_i. Then the estimated pdf at x_i is $\wp_{\text{x,est}}(x_i) = \Delta N_i/(N_{\text{tot}}\Delta x)$.* (5.2)

For example, if we want to find the pdf of adult human heights, we'll get a fairly continuous distribution if we take Δx to be about 1 cm or less, and N_{tot} large enough to have many samples in each bin in the range of interest. Notice that Equation 5.1 implies that[3]

> *A probability density function for x has dimensions inverse to those of x.* (5.3)

---

[2] Similarly, in Figure 3.1 (page 37), the larger number of bins in panel (b) means that each bar is shorter than in (a).

[3] Just as mass density (kilograms per cubic meter) has different units from mass (kilograms), so the terms "probability density" here, and "probability mass" in Section 3.3.1, were chosen to emphasize the different units of these quantities. Many authors simply use "probability distribution" for either the discrete or continuous case.

We can also express a continuous pdf in the language of events:[4] Let $\mathsf{E}_{x_0,\Delta x}$ be the event containing all outcomes for which the value of x lies within a range of width Δx around the value x_0. Then Equation 5.1 says that

$$\wp(x_0) = \lim_{\Delta x \to 0} \big(\mathcal{P}(\mathsf{E}_{x_0,\Delta x})/(\Delta x)\big). \quad \text{probability density function} \tag{5.4}$$

$\wp_x(x_0)$ is not the probability to observe a particular value for x; as mentioned earlier, that's always zero. But once we know $\wp(x)$, then the probability that a measurement will fall into a finite *range* is $\int_{x_1}^{x_2} \mathrm{d}x\, \wp(x)$. Thus, the normalization condition, Equation 3.4 (page 42), becomes

$$\int \mathrm{d}x\, \wp(x) = 1, \quad \textbf{normalization condition}, \text{ continuous case} \tag{5.5}$$

where the integral runs over all allowed values of x. That is, the area under the curve defined by $\wp(x)$ must always equal 1. As in the discrete case, a pdf is always nonnegative. *Unlike* the discrete case, however, a pdf need not be everywhere smaller than 1: It can have a high, but narrow, spike and still obey Equation 5.5.

T2 *Section 5.2.1′ (page 114) discusses an alternative definition of the pdf used in mathematical literature.*

5.2.2 Three key examples: Uniform, Gaussian, and Cauchy distributions

Uniform, continuous distribution

Consider a probability density function that is *constant* throughout the range $x_{\min}$ to $x_{\max}$:

$$\wp_{\text{unif}}(x) = \begin{cases} 1/(x_{\max} - x_{\min}) & \text{if } x_{\min} \le x \le x_{\max}; \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

The formula resembles the discrete case,[5] but note that now $\wp_{\text{unif}}(x)$ will have dimensions, if the variable x does.

Gaussian distribution

The famous "bell curve" is actually a family of functions defined by the formula

$$f(x; \mu_x, \sigma) = A\mathrm{e}^{-(x-\mu_x)^2/(2\sigma^2)}, \tag{5.7}$$

where x ranges from $-\infty$ to $+\infty$. Here A and σ are positive constants; μ_x is another constant.

[4] See Section 3.3.1 (page 41).
[5] See Section 3.3.2 (page 43).

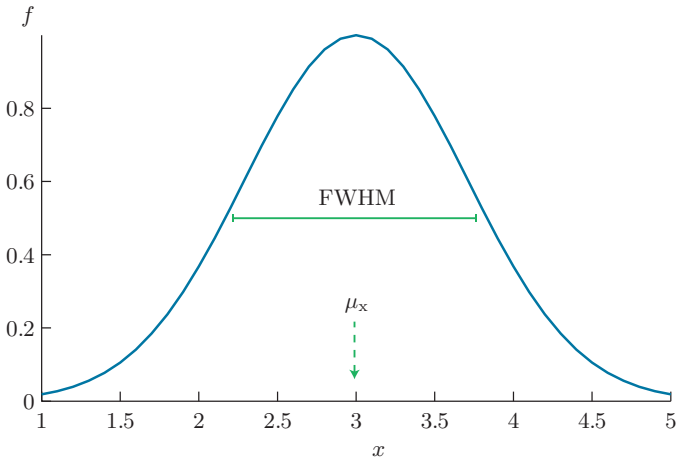


Figure 5.1 [Mathematical function.] **The function defined in Equation 5.7,** with $A = 1$, $\mu_x = 3$, and $\sigma = 1/\sqrt{2}$. Although the function is very small outside the range shown, it is nonzero for *any* x. The abbreviation FWHM refers to the full width of this curve at one half the maximum value, which in this case equals $2\sqrt{\ln 2}$. The Gaussian distribution $\wp_{\text{gauss}}(x; 3, 1/\sqrt{2})$ equals this f times $1/\sqrt{\pi}$ (see Equation 5.8).

Figure 5.1 shows an example of this function. Graphing it for yourself, and playing with the parameters, is a good way to bring home the point that the bell curve is a bump function centered at μ_x (that is, it attains its maximum there) with *width* controlled by the parameter σ. Increasing the value of σ makes the bump wider.

The function f in Equation 5.7 is everywhere nonnegative, but this is not enough: It's only a candidate for a probability density function if it also satisfies the normalization condition, Equation 5.5. Thus, the constant A appearing in it isn't free; it's determined in terms of the other parameters by

$$1/A = \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-(x-\mu_x)^2/(2\sigma^2)}.$$

Even if you don't have your computer handy, you can make some progress evaluating this integral. Changing variables to $y = (x - \mu_x)/(\sigma\sqrt{2})$ converts it to

$$1/A = \sigma\sqrt{2} \int_{-\infty}^{\infty} \mathrm{d}y\, \mathrm{e}^{-y^2}.$$

At this point we are essentially done: We have extracted all the dependence of A on the other parameters (that is, $A \propto \sigma^{-1}$). The remaining integral is just a universal constant, which we could compute just once, or look up. In fact, it equals $\sqrt{\pi}$. Substituting into Equation 5.7 yields "the" Gaussian distribution, or rather a family of distributions defined by the probability density functions[6]

$$\wp_{\text{gauss}}(x; \mu_x, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(x-\mu_x)^2/(2\sigma^2)}. \quad \textbf{Gaussian distribution} \tag{5.8}$$

[6] The special case $\mu_x = 0, \sigma = 1$ is also called the **normal distribution**.

The appearance of $1/\sigma$ in front of the exponential has a simple interpretation. Decreasing the value of σ makes the exponential function more narrowly peaked. In order to maintain fixed area under the curve, we must therefore make the curve taller; the factor $1/\sigma$ accomplishes this. This factor also gives $\wp(x)$ the required dimensions (inverse to those of x).

The Gaussian distribution has arrived with little motivation, merely a nontrivial example of a continuous distribution on which to practice some skills. We'll understand its popularity a bit later, when we see how it emerges in a wide class of real situations. First, though, we introduce a counterpoint, another similar-looking distribution with some surprising features.

Cauchy distribution

Consider the family of probability density functions of the form[7]

$$\wp_{\text{cauchy}}(x;\mu_x,\eta) = \frac{A}{1+\left(\frac{x-\mu_x}{\eta}\right)^2}. \quad \textbf{Cauchy distribution} \tag{5.9}$$

Here, as before, μ_x is a parameter specifying the most probable value of x (that is, it specifies the distribution's center). η is a constant a bit like σ in the Gaussian distribution; it determines how wide the bump is.

Your Turn 5A

a. Find the required value of the constant A in Equation 5.9 in terms of the other constants, using a method similar to the one that led to Equation 5.8. Graph the resulting pdf, and compare with a Gaussian having the same FWHM.
b. Your graph may seem to say that there isn't much difference between the Gaussian and Cauchy pdfs. To see the huge difference more clearly, plot them together on semilog axes (logarithmic axis for the $\wp$, linear for x), and compare them again.

Section 5.4 will discuss real situations in which Cauchy, and related, distributions arise.

5.2.3 Joint distributions of continuous random variables

Just as in the discrete case, we will often be interested in joint distributions, that is, in random systems whose outcomes are sets of two or more continuous values (see Section 3.4.2). The same reasoning that led to the definition of the pdf (Equation 5.1) then leads us to define ΔN as the number of observations for which x lies in a particular range around x_0 of width Δx, *and* y also lies in a particular range of width Δy, and so on. To get a good limit, then, we must divide ΔN by the product $(\Delta x)(\Delta y)\cdots$. Equivalently, we can imitate Equation 5.4:

$$\wp(x_0,y_0) = \lim_{\Delta x,\Delta y\to 0}\left(\mathcal{P}(\mathsf{E}_{x_0,\Delta x}\ \textbf{and}\ \mathsf{E}_{y_0,\Delta y})/(\Delta x\,\Delta y)\right).$$

[7] Some authors call them Lorentzian or Breit-Wigner distributions.

Your Turn 5B

Find appropriate generalizations of the dimensions and normalization condition (Idea 5.3 and Equation 5.5) for the case of a continuous, joint distribution.

We can also extend the notion of conditional probability (see Equation 3.10, page 45):

$$\wp(x \mid y) = \wp(x, y)/\wp(y). \tag{5.10}$$

Thus,

> *The dimensions of the conditional pdf $\wp(x \mid y)$ are inverse to those of x, regardless of the dimensions of y.*

Example Write a version of the Bayes formula for $\wp(x \mid y)$, and verify that the units work out properly.

Solution Begin with a formula similar to Equation 5.10 but with x and y quantities reversed. Comparing the two expressions and imitating Section 3.4.4 (page 52) yields

$$\wp(y \mid x) = \wp(x \mid y)\wp(y)/\wp(x). \tag{5.11}$$

On the right-hand side, $\wp(x)$ in the denominator cancels the units of $\wp(x \mid y)$ in the numerator. Then the remaining factor $\wp(y)$ gives the right-hand side the appropriate units to match the left-hand side.

The continuous form of the Bayes formula will prove useful in the next chapter, providing the starting point for localization microscopy. You can similarly work out a version of the formula for the case when one variable is discrete and the other is continuous.

5.2.4 Expectation and variance of the example distributions

Continuous distributions have descriptors similar to the discrete case. For example, the expectation is defined by[8]

$$\langle f \rangle = \int \mathrm{d}x\, f(x)\wp(x).$$

Note that $\langle f \rangle$ has the same dimensions as f, because the units of $\mathrm{d}x$ cancel those of $\wp(x)$.[9] The variance of f is defined by the same formula as before, Equation 3.20 (page 55); thus it has the same dimensions as f^2.

Your Turn 5C

a. Find $\langle x \rangle$ for the Uniform continuous distribution on some range $a < x < b$. Repeat for the pdf $\wp_{\text{gauss}}(x; \mu_x, \sigma)$.
b. Find var x for the Uniform continuous distribution.

[8] Compare the discrete version Equation 3.19 (page 53).
[9] The same remark explains how the normalization integral (Equation 5.5) can equal the pure number 1.

Your Turn 5D

The Gaussian distribution has the property that its expectation and most probable value are equal. Think: What sort of distribution could give *unequal* values?

The variance of a Gaussian distribution is a bit more tricky; let's first guess its general form. The spread of a distribution is unchanged if we just shift it.[10] Changing μ_x just shifts the Gaussian, so we don't expect μ_x to enter into the formula for the variance. The only other relevant parameter is σ. Dimensional analysis shows that the variance must be a constant times σ^2.

To be more specific than this, we must compute the expectation of x^2. We can employ a trick that we've used before:[11] Define a function $I(b)$ by

$$I(b) = \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-bx^2}.$$

Section 5.2.2 explained how to evaluate this normalization-type integral; the result is $I(b) = \sqrt{\pi/b}$. Now consider the derivative $\mathrm{d}I/\mathrm{d}b$. On one hand, it's

$$\mathrm{d}I/\mathrm{d}b = -(1/2)\sqrt{\pi/b^3}. \tag{5.12}$$

But also,

$$\mathrm{d}I/\mathrm{d}b = \int_{-\infty}^{\infty} \mathrm{d}x \frac{\mathrm{d}}{\mathrm{d}b}\mathrm{e}^{-bx^2} = -\int_{-\infty}^{\infty} \mathrm{d}x\, x^2 \mathrm{e}^{-bx^2}. \tag{5.13}$$

That last integral is the one we need in order to compute $\langle x^2 \rangle$. Setting the right sides of Equations 5.12 and 5.13 equal to each other and evaluating at $b = (2\sigma^2)^{-1}$, gives

$$\int_{-\infty}^{\infty} \mathrm{d}x\, x^2 \mathrm{e}^{-x^2/(2\sigma^2)} = \tfrac{1}{2}\pi^{1/2}(2\sigma^2)^{3/2}.$$

With this preliminary result, we can finally evaluate the variance of a Gaussian distribution centered on zero:

$$\operatorname{var} x = \langle x^2 \rangle = \int \mathrm{d}x\, \wp_{\text{gauss}}(x; 0, \sigma) x^2 = \left[(2\pi\sigma^2)^{-1/2}\right]\left[\tfrac{1}{2}\pi^{1/2}(2\sigma^2)^{3/2}\right] = \sigma^2. \tag{5.14}$$

Because the variance doesn't depend on where the distribution is centered, we conclude more generally that

$$\operatorname{var} x = \sigma^2 \quad \text{if } x \text{ is drawn from } \wp_{\text{gauss}}(x; \mu_x, \sigma). \tag{5.15}$$

Example Find the variance of the Cauchy distribution.

Solution Consider the Cauchy distribution centered on zero, with $\eta = 1$. This time, the integral that defines the variance is

[10] See Your Turn 3L (page 56).
[11] See the Example on page 77.

$$\int_{-\infty}^{\infty} dx \, \frac{x^2}{\pi} \frac{1}{1+x^2}.$$

This integral is infinite, because at large $|x|$ the integrand approaches a constant.

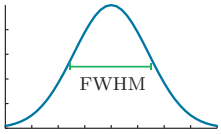


Figure 5.1 (page 100)

Despite this surprising result, the Cauchy distribution is normalizable, and hence it's a perfectly legitimate probability density function. The problem lies not with the distribution, but with the choice of variance as a descriptor: The variance is very sensitive to outliers, and a Cauchy distribution has many more of these than does a Gaussian.

Other descriptors of spread work just fine for the Cauchy distribution, however. For example, we can use full width at half maximum (FWHM; see Figure 5.1[12]) instead of variance to describe its spread.

T2 *Section 5.2.4′ (page 114) introduces another measure of spread that is useful for long-tail distributions: the interquartile range.*

5.2.5 Transformation of a probability density function

The definition of probability density function creates an important difference from the discrete case. Suppose that you have recorded many vocalizations of some animal, perhaps a species of whale. The intensity and pitch vary over time. You'd like to characterize these sounds, perhaps to see how they vary with species, season, and so on. One way to begin might be to define x as the intensity of sound emitted (in watts, abbreviated W) and create an estimated pdf $\wp_x$ from many observations of x. A colleague, however, may believe that it's more meaningful to report the related quantity $y = 10 \log_{10}(x/(1\,\mathrm{W}))$, the sound intensity on a "decibel" scale. That colleague will then report the pdf $\wp_y$.

To compare your results, you need to *transform* your result from your choice of variable x to your colleague's choice y. To understand transformation in general, suppose that x is a continuous random variable with some known pdf $\wp_x(x)$. If we collect a large number of draws from that distribution ("measurements of x"), the fraction that lie between $x_0 - \frac{1}{2}\Delta x$ and $x_0 + \frac{1}{2}\Delta x$ will be $\wp_x(x_0)\Delta x$.[13] Now define a new random variable y to be some function applied to x, or $y = G(x)$. This y is not independent of x; it's just another description of the same random quantity reported by x. Suppose that G is a strictly increasing or decreasing function—a **monotonic** function.[14] In the acoustic example above, $G(x) = 10 \log_{10}(x/1\,\mathrm{W})$ is a strictly increasing function (see Figure 5.2); thus, its derivative dG/dx is everywhere positive.

To find $\wp_y$ at some point y_0, we now ask, for a small interval Δy: How often does y lie within a range $\pm\frac{1}{2}\Delta y$ of y_0? Figure 5.2 shows that, if we choose the y interval to be the image of the x interval, then the *same fraction* of all the points lie in this interval in either description. We know that $y_0 = G(x_0)$. Also, because Δx is small, Taylor's theorem gives $\Delta y \approx (\Delta x)(dG/dx|_{x_0})$, and so

$$\left[\wp_y(G(x_0))\right]\left[(\Delta x)\left.\frac{dG}{dx}\right|_{x_0}\right] = \wp_x(x_0)(\Delta x).$$

[12] See also Problem 5.10.

[13] See Equation 5.1 (page 98).

[14] T2 Thus, G associates exactly one x value to each y in its range. If G is not monotonic, the notation gets more awkward but we can still get a result analogous to Equation 5.16.

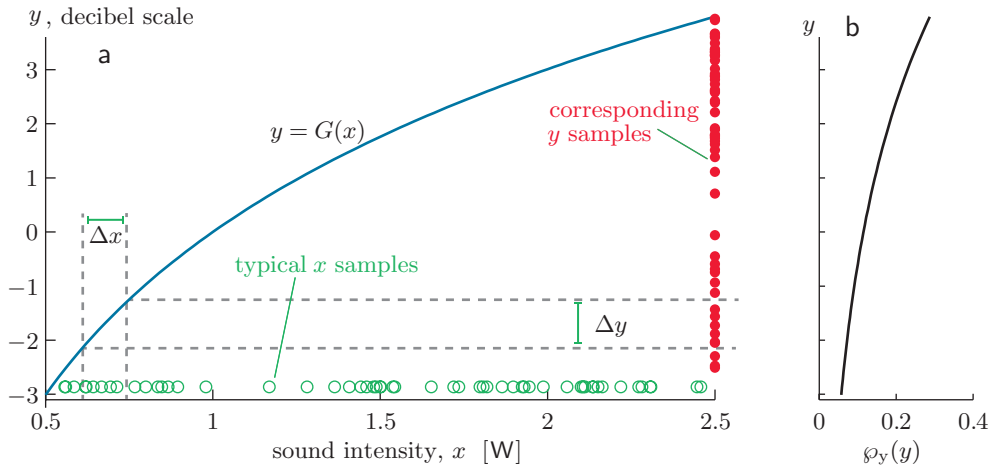


Figure 5.2 [Mathematical functions.] **Transformation of a pdf.** (a) *Open circles* on the horizontal axis give a cloud representation of a Uniform distribution $\wp_x(x)$. These representative samples are mapped to the vertical axis by the function $G(x) = 10\log_{10}(x/(1\,\mathrm{W}))$ and are shown there as *solid circles*. They give a cloud representation of the transformed distribution $\wp_y(y)$. One particular interval of width Δx is shown, along with its transformed version on the y axis. Both representations agree that this bin contains five samples, but they assign it different widths. (b) The transformed pdf (horizontal axis), determined by using Equation 5.16, reflects the non-Uniform density of the solid circles in (a).

Dividing both sides by $(\Delta x)(\mathrm{d}G/\mathrm{d}x|_{x_0})$ gives the desired formula for $\wp_y$:

$$\wp_y(y_0) = \wp_x(x_0) \Big/ \left.\frac{\mathrm{d}G}{\mathrm{d}x}\right|_{x_0} \qquad \text{for monotonically increasing } G. \tag{5.16}$$

The right side of this formula is a function of y_0, because we're evaluating it at $x_0 = G^{-1}(y_0)$, where G^{-1} is the inverse function to G.

Your Turn 5E

a. Think about how the dimensions work in Equation 5.16, for example, in the situation where x has dimensions $\mathbb{L}$ and $G(x) = x^3$. Your answer provides a useful mnemonic device for the formula.
b. Why aren't the considerations of this section needed when we study *discrete* probability distributions?

Example Go through the above logic again for the case of a function G that's monotonically *decreasing*, and make any necessary changes.

Solution In this case, the width of the y interval corresponding to Δx is $-\Delta x\big(\mathrm{d}G/\mathrm{d}x\big)$, a positive quantity. Using the absolute value covers both cases:

$$\wp_y(y_0) = \wp_x(x_0) \Big/ \left|\left.\frac{\mathrm{d}G}{\mathrm{d}x}\right|_{x_0}\right|. \qquad \text{transformation of a pdf, where } x_0 = G^{-1}(y_0) \tag{5.17}$$

The transformation formula just found will have repercussions when we discuss model selection in Section 6.2.3.

5.2.6 Computer simulation

The previous section stressed the utility of transformations when we need to convert a result from one description to another. We now turn to a second practical application, simulating draws from a specified distribution by using a computer. Chapter 8 will use these ideas to create simulations of cell reaction networks.

Equation 5.17 has an important special case: If y is the *Uniformly* distributed random variable on the range $[0, 1]$, then $\wp_Y(y) = 1$ and $\wp_X(x_0) = \left|dG/dx|_{x_0}\right|$. This observation is useful when we wish to simulate a random system with some arbitrarily specified probability density function:

> *To simulate a random system with a specified pdf $\wp_X$, find a function G whose derivative equals $\pm\wp_X$ and that maps the desired range of x onto the interval $[0, 1]$. Then apply the inverse of G to a Uniformly distributed variable y; the resulting x values will have the desired distribution.* (5.18)

Example The probability density function $\wp(x) = e^{-x}$, where x lies between zero and infinity, will be important in later chapters. Apply Idea 5.18 to simulate draws from a random variable with this distribution.

Solution To generate x values, we need a function G that solves $|dG/dx| = e^{-x}$. Thus,

$$G(x) = \text{const} \pm e^{-x}.$$

Applying functions of this sort to the range $[0, \infty)$, we see that the choice e^{-x} works. The inverse of that function is $x = -\ln y$.

Try applying $-\ln$ to your computer's random number generator, and making a histogram of the results.

Your Turn 5F

Think about the discussion in Section 4.2.5 (page 73) of how to get a computer to draw from a specified *discrete* distribution (for example, the Poisson distribution). Make a connection to the above discussion.

Your Turn 5G

Apply Idea 5.18 to the Cauchy distribution, Equation 5.9 (page 101), with $\mu_X = 0$ and $\eta = 1$. Use a computer to generate some draws from your distribution, histogram them, and confirm that they have the desired distribution.

5.3 More About the Gaussian Distribution

5.3.1 The Gaussian distribution arises as a limit of Binomial

The Binomial distribution is very useful, but it has two unknown parameters: the number of draws M and the probability ξ to flip heads. Section 4.3 described a limiting case, in which

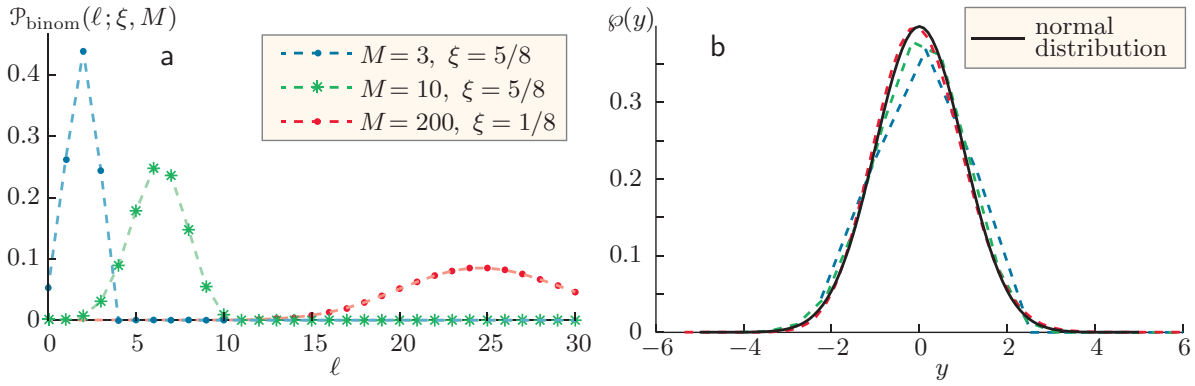


Figure 5.3 [Mathematical functions.] **The Gaussian distribution as a limit.** (a) Three examples of Binomial distributions. (b) The same three discrete distributions as in (a) have been modified as described in the text. In particular, each curve has been rescaled: Instead of all the points summing to 1, the scale factor $1/\Delta y$ has been applied to ensure that the *area* under each curve equals 1. For comparison, the *solid curve* shows the Gaussian distribution $\wp_{\text{gauss}}$ with $\mu_y = 0$ and $\sigma = 1$. The figure shows that $M = 3$ or 10 give only roughly Gaussian-shaped distributions but that distributions with large M and $M\xi$ are very nearly Gaussian. See also Problem 5.14.

the Binomial distribution "forgets" the individual values of these parameters, "remembering" only their product $\mu = M\xi$. The appropriate limit was large M at fixed μ.

You may already have noticed, however, an even greater degree of universality when μ is also large.[15] Figure 5.3a shows three examples of Binomial distributions. When both M and $M\xi$ are large, the curves become smooth and symmetric, and begin to look very similar to Gaussian distributions.

It's true that the various Binomial distributions all differ in their expectations and variances. But these superficial differences can be eliminated by changing our choice of variable, as follows: First, let

$$\mu_\ell = M\xi \quad \text{and} \quad s = \sqrt{M\xi(1-\xi)}.$$

Then define the new random variable $y = (\ell - \mu_\ell)/s$. Thus, y always has the same expectation, $\langle y \rangle = 0$, and variance, var $y = 1$, regardless of what values we choose for M and ξ.

We'd now like to compare other features of the y distribution for different values of M and ξ, but first we face the problem that the list of allowed values (the sample space) for y depends on the parameters. For example, the spacing between successive discrete values is $\Delta y = 1/s$. But instead of a direct comparison, we can divide the discrete distribution $\mathcal{P}(y; M, \xi)$ by Δy. If we then take the limit of large M, we obtain a family of probability density functions, each for a continuous random variable y in the range $-\infty < y < \infty$. It does make sense to compare these pdfs for different values of ξ, and remarkably[16]

> *For any fixed value of ξ, the distribution of y approaches a universal form. That is, it "forgets" the values of both M and ξ, as long as M is large enough. The universal limiting pdf is Gaussian.*

[15] See Problem 4.6.

[16] T2 In Problem 5.14 you'll prove a more precise version of this claim.

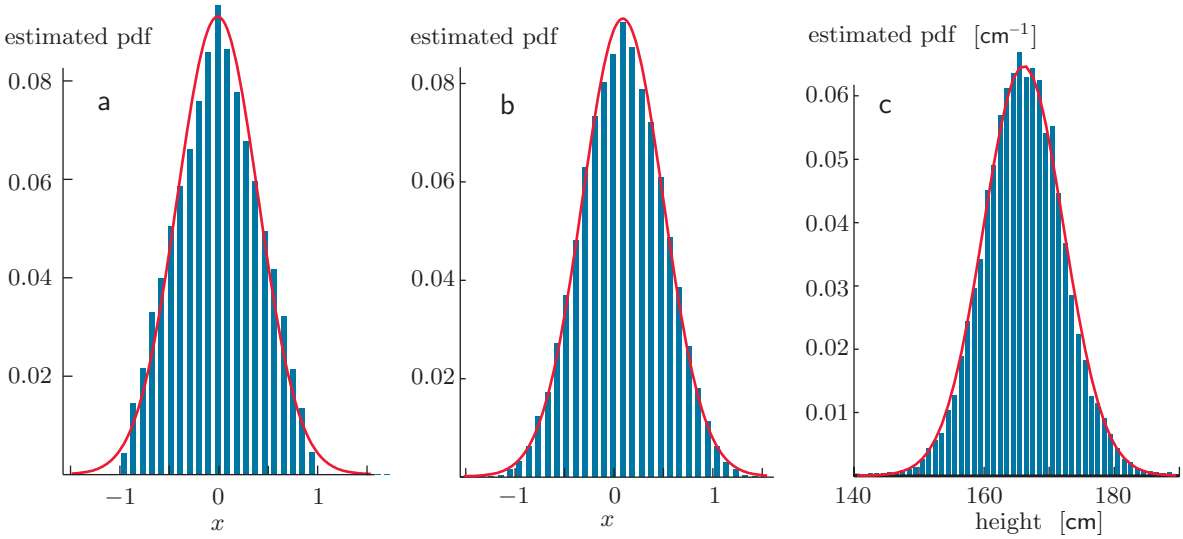


Figure 5.4 [Computer simulations.] **The central limit theorem at work.** (a) *Bars:* Histogram of 50 000 draws of a random variable defined as the sum of two independent random variables, each Uniformly distributed on the range $[-1/2, +1/2]$. The *curve* shows a Gaussian distribution with the same expectation and variance for comparison. (b) *Bars:* 50 000 draws from the sum of four such variables, scaled by $1/\sqrt{2}$ to give the same variance as in (a). The *curve* is the same as in (a). (c) [Empirical data with fit.] Distribution of the heights of a similarly large sample of 21-year-old men born in southern Spain. The curve is the best-fitting Gaussian distribution. [Data from María-Dolores & Martínez-Carrión, 2011.]

Figure 5.3b illustrates this result. For example, the figure shows that even a highly asymmetric Bernoulli trial, like $\xi = 1/8$, gives rise to the symmetric Gaussian for large enough M.

5.3.2 The central limit theorem explains the ubiquity of Gaussian distributions

The preceding subsection began to show why Gaussian distributions arise frequently: Many interesting quantities really can be regarded as sums of many independent Bernoulli trials, for example, the number of molecules of a particular type captured in a sample drawn from a well-mixed solution. And in fact, the phenomenon noted in Section 5.3.1 is just the beginning. Here is another example of the same idea at work.

Let $\wp_1(x)$ denote the continuous Uniform distribution on the range $-1/2 < x < 1/2$. Its probability density function does not look very much like any Gaussian, not even one chosen to have the same expectation and variance. Nevertheless, Figure 5.4a shows that the *sum of two* independent random variables, each drawn from $\wp_1$, has a distribution that looks a bit more like a Gaussian, although unlike a Gaussian, it equals zero outside a finite range. And the sum of *just four* such variables looks very much like a Gaussian (Figure 5.4b).[17]

This observation illustrates a key result of probability theory, the **central limit theorem**. It applies when we have M independent random variables (continuous or discrete), each

[17] See Problem 5.7. Incidentally, this exercise also illustrates the counterpoint between analytic and simulation methods. The distribution of a sum of random variables is the convolution of their individual distributions (Section 4.3.5). It would be tedious to work out an exact formula for the convolution of even a simple distribution with itself, say 10 times. But it's easy to make sets of 10 draws from that distribution, add them, and histogram the result.

drawn from identical distributions. The theorem states roughly that, for large enough M, the quantity $x_1 + \cdots + x_M$ is always distributed as a Gaussian. We have discussed examples where x was Bernoulli or Uniform, but actually the theorem holds *regardless* of the distribution of the original variable, as long as it has finite expectation and variance.

5.3.3 When to use/not use a Gaussian

Indeed, in Nature we often do observe quantities that reflect the additive effects of many independent random influences. For example, human height is a complex phenotypic trait, dependent on hundreds of different genes, each of which is dealt to us at least partially independently of the others. It's reasonable to suppose that these genes have at least partially additive effects on overall height,[18] and that the ones making the biggest contributions are roughly equal in importance. In such a situation, we may expect to get a Gaussian distribution, and indeed many phenotypic traits, including human height, do follow this expectation (Figure 5.4c).[19]

Your Turn 5H

Problem 4.5 introduced a model for molecular diffusion based on trajectories that take steps of $\pm d$ in each direction. But surely this is an oversimplification: The minute kicks suffered by a suspended particle must have a variety of strengths, and so must result in displacements by a variety of distances. Under what circumstances may we nevertheless expect the random walk to be a good model of diffusive motion?

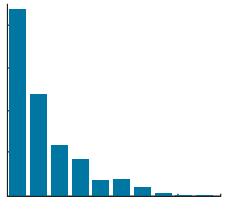


Figure 3.2b (page 38)

In our lab work we sometimes measure the same quantity independently several times, then take the average of our measurements. The central limit theorem tells us why in these situations we generally see a Gaussian distribution of results. However, we should observe some caveats:

- Some random quantities are *not* sums of many independent, identically distributed random variables. For example, the blip waiting times shown in Figure 3.2b are far from being Gaussian distributed. Unlike a Gaussian, their distribution is very asymmetrical, reaching its maximum at its extreme low end.
- Even if a quantity does seem to be such a sum, and its distribution does appear fairly close to Gaussian near its peak, nevertheless for any finite N there may be significant discrepancies in the tail region (see Figure 5.4a), and for some applications such low-probability events may be important. Also, many kinds of experimental data, such as the *number* of blips in a time window, are Poisson distributed; for low values of μ such distributions can also be far from Gaussian.
- An observable quantity may indeed be the sum of contributions from many sources, but they may be interdependent. For example, spontaneous neural activity in the brain involves the electrical activity of many nerve cells (**neurons**), each of which is connected to many others. When a few neurons fire, they may tend to trigger others, in an "avalanche" of activity. If we add up all the electrical activity in a region of the brain, we will see a signal with peaks reflecting the total numbers of neurons firing in each event. These event magnitudes were found to have a distribution that was far from Gaussian (Figure 5.5).

[18] That is, we are neglecting the possibility of nonadditive gene interaction, or **epistasis**.

[19] Height also depends on nutrition and other environmental factors. Here, we are considering a roughly homogeneous population and supposing that any remaining variation in environment can also be roughly modeled as additive, independent random factors.

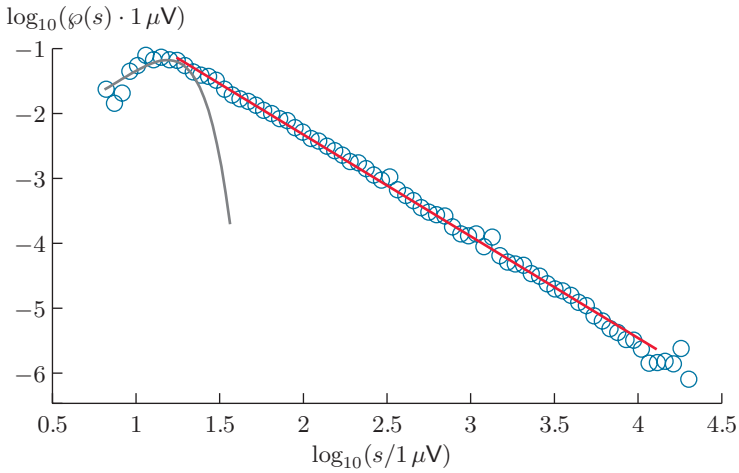


Figure 5.5 [Experimental data with fit.] **Power-law distribution of neural activity.** Slices of brain tissue from rats were cultured on arrays of extracellular electrodes that recorded the neurons' spontaneous activities. The electric potential outside the cells was found to have events consisting of bursts of activity from many cells. The total measured activity s in each burst (the "magnitude" of the event) was tabulated and used to find an estimated pdf. This log-log plot shows that the distribution has a power-law form (*red line*), with exponent −1.6. For comparison, the *gray line* shows an attempted fit to a Gaussian. Similar results were observed in intact brains of live animals. [Data from Gireesh & Plenz, 2008.]

- We have already seen that some distributions have *infinite* variance. In such cases, the central limit theorem does not apply, even if the quantity in question is a sum of independent contributions.[20]

5.4 More on Long-tail Distributions

The preceding section pointed out that not every measurement whose distribution seems to be bump-shaped will actually be Gaussian. For example, the Gaussian distribution far from its center falls as a constant times $\exp(-x^2/(2\sigma^2))$, whereas the Cauchy distribution,[21] which looks superficially similar, approaches a constant times $x^{-\alpha}$, with $\alpha = 2$. More generally, there are many random systems with **power-law distributions**; that is, they exhibit this kind of limiting behavior for some constant α. Power-law distributions are another example of the long-tail phenomenon mentioned earlier, because any power of x falls more slowly at large x than the Gaussian function.[22]

To see whether an empirically obtained distribution is of power-law form, we can make a **log-log plot**, so named because both the height of a point and its horizontal position are proportional to the logarithms of its x and y values. The logarithm of $Ax^{-\alpha}$ is $\log(A) - \alpha \log(x)$, which is a linear function of $\log x$. Thus, this function will appear on a log-log plot as a straight line,[23] with slope $-\alpha$. A power-law distribution will therefore have this straight-line form for large enough x. Figure 5.5 shows a biological example of such a distribution.

[20] See Problem 5.13.
[21] See Equation 5.9 (page 101).
[22] See Section 4.4.2 (page 81). You'll investigate the weight in the tails of various distributions in Problem 5.10.
[23] See Problem 5.11.

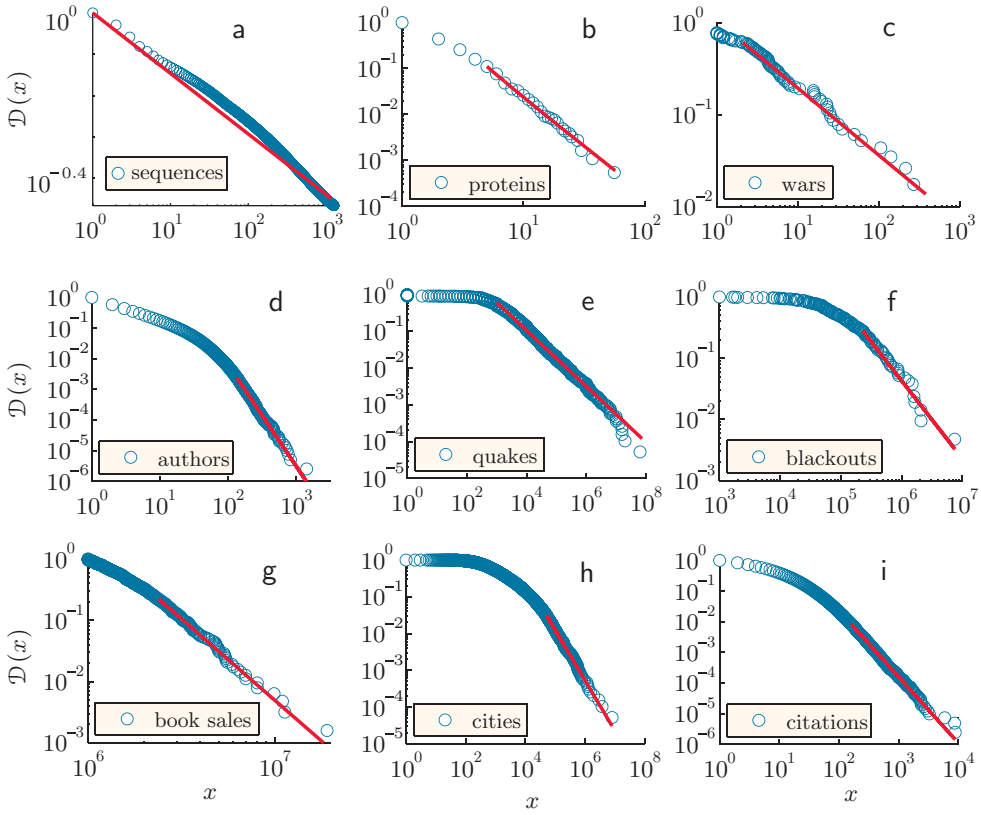


Figure 5.6 [Empirical data with fits.] **Power law distributions in many contexts.** Each panel shows the complementary cumulative distribution function $\mathcal{D}(x)$, Equation 5.19 (or its discrete analog), and a power-law fit, for a different dataset. (a) The probabilities of occurrence of various antibody sequences in the immune system of a single organism versus rank order x. The entire antibody repertoire of a zebrafish was sequenced, and a list was made of sequences of the "D region" of each antibody. Over a wide range, the probability followed the approximate form $\wp(x) \propto x^{-\alpha}$ with $\alpha \approx 1.15$. (b) The numbers of distinct interaction partners of proteins in the protein-interaction network of the yeast *S. cerevisiae* ($\alpha = 3$). (c) The relative magnitudes of wars from 1816 to 1980, that is, the number of battle deaths per 10 000 of the combined populations of the warring nations ($\alpha = 1.7$). (d) The numbers of authors on published academic papers in mathematics ($\alpha = 4.3$). (e) The intensities of earthquakes occurring in California between 1910 and 1992 ($\alpha = 1.6$). (f) The magnitudes of power failures (number of customers affected) in the United States ($\alpha = 2.3$). (g) The sales volumes of bestselling books in the United States ($\alpha = 3.7$). (h) The populations of cities in the United States ($\alpha = 2.4$). (i) The numbers of citations received by published academic papers ($\alpha = 3.2$). [Data from Clauset et al., 2009, and Mora et al., 2010.]

Equivalently, we can examine a related quantity called the **complementary cumulative distribution**,[24] the probability of drawing a value of x larger than some specified value:

$$\mathcal{D}(x) = \int_x^\infty \mathrm{d}x' \, \wp(x'). \tag{5.19}$$

[24]The qualifier "complementary" distinguishes $\mathcal{D}$ from a similar definition with the integral running from $-\infty$ to x.

$\mathcal{D}(x)$ is always a decreasing function. For the case of a power-law distribution, the log-log graph of $\mathcal{D}(x)$ is moreover a straight line.[25] Figure 5.6 shows that, remarkably, pdfs of approximately power-law form arise in various natural phenomena, and even human society.

T_2 *Section 5.4′ (page 115) discusses another example of a power-law distribution.*

THE BIG PICTURE

As in earlier chapters, this one has focused on a small set of illustrative probability distributions. The ones we have chosen, however, turn out to be useful for describing a remarkable range of biological phenomena. In some cases, this is because distributions like the Gaussian arise in contexts involving many independent actors (or groups of actors), and such situations are common in both the living and nonliving worlds.[26] Other distributions, including power laws, are observed in large systems with more complex interactions.

Chapter 6 will apply the ideas developed in preceding chapters to our program of understanding *inference,* the problem of extracting conclusions from partially random data.

KEY FORMULAS

- *Probability density function (pdf) of a continuous random variable:*

$$\wp_{\mathrm{x}}(x_0) = \lim_{\Delta x \to 0} \left(\lim_{N_{\mathrm{tot}} \to \infty} \frac{\Delta N}{N_{\mathrm{tot}} \Delta x} \right) = \lim_{\Delta x \to 0} \left(\mathcal{P}(\mathsf{E}_{x_0, \Delta x}) / (\Delta x) \right). \qquad (5.1) + (5.4)$$

  Note that $\wp_{\mathrm{x}}$ has dimensions inverse to those of x. The subscript "x" can be omitted if this does not cause confusion. Joint, marginal, and conditional distributions are defined similarly to the discrete case.
- *Estimating a pdf:* Given some observations of x, choose a set of bins that are wide enough to each contain many observations. Find the frequencies ΔN_i for each bin centered on x_i. Then the estimated pdf at x_i is $\Delta N_i / (N_{\mathrm{tot}} \Delta x)$.
- *Normalization and moments of continuous distribution:* $\int \mathrm{d}x \, \wp(x) = 1$. The expectation and variance of a function of x are then defined analogously to discrete distributions, for example, $\langle f \rangle = \int \mathrm{d}x \, \wp(x) f(x)$.
- *Continuous version of Bayes formula:*

$$\wp(y \mid x) = \wp(x \mid y) \wp(y) / \wp(x). \qquad (5.11)$$

- *Gaussian distribution:* $\wp_{\mathrm{gauss}}(x; \mu_{\mathrm{x}}, \sigma) = (\sigma \sqrt{2\pi})^{-1} \exp(-(x - \mu_{\mathrm{x}})^2 / (2\sigma^2))$. The random variable x, and the parameters μ_{x} and σ, can have any units, but they must all match. $\langle x \rangle = \mu_{\mathrm{x}}$ and $\mathrm{var}\, x = \sigma^2$.
- *Cauchy distribution:*

$$\wp_{\mathrm{cauchy}}(x; \mu_{\mathrm{x}}, \eta) = \frac{A}{1 + \left(\frac{x - \mu_{\mathrm{x}}}{\eta} \right)^2}.$$

[25] In Problem 5.11 you'll contrast the corresponding behavior in the case of a Gaussian distribution.

[26] The Exponential distribution to be studied in Chapter 7 has this character as well, and enjoys a similarly wide range of application.

This is an example of a power-law distribution, because $\wp_{\text{cauchy}} \to A\eta^2 x^{-2}$ at large $|x|$. The constant A has a specific relation to η; see Your Turn 5A (page 101). The random variable x, and the parameters μ_x and η, can have any units, but they must all match.

- *Transformation of a pdf:* Suppose that x is a continuous random variable with probability density function $\wp_x$. Let y be a new random variable, defined by drawing a sample x and applying a strictly increasing function G. Then $\wp_y(y_0) = \wp_x(x_0)/G'(x_0)$, where $y_0 = G(x_0)$ and $G' = \mathrm{d}G/\mathrm{d}x$. (One way to remember this formula is to recall that it must be valid even if x and y have different dimensions.) If G is strictly *de*creasing we get a similar formula but with $|\mathrm{d}G/\mathrm{d}x|$.

FURTHER READING

Semipopular:
Hand, 2008.

Intermediate:
Bolker, 2008; Denny & Gaines, 2000; Otto & Day, 2007, §P3; Ross, 2010.

Technical:
Power-law distributions in many contexts: Clauset et al., 2009.
Power-law distribution in neural activity: Beggs & Plenz, 2003.
Complete antibody repertoire of an animal: Weinstein et al., 2009.

T_2 **Track 2**

5.2.1′ Notation used in mathematical literature

The more complex a problem, the more elaborate our mathematical notation must be to avoid confusion and even outright error. But conversely, elaborate notation can unnecessarily obscure less complex problems; it helps to be flexible about our level of precision. This book generally uses a notation that is customary in physics literature and is adequate for many purposes. But other books use a more formal notation, and a word here may help the reader bridge to those works.

For example, we have been a bit imprecise about the distinction between a random variable and the specific values it may take. Recall that Section 3.3.2 defined a random variable to be a function on sample space. Every "measurement" generates a point of sample space, and evaluating the function at that point yields a numerical value. To make this distinction clearer, some authors reserve capital letters for random variables and lowercase for possible values.

Suppose that we have a discrete sample space, a random variable X, and a number x. Then the event $\mathsf{E}_{X=x}$ contains the outcomes for which X took the specific value x, and $\mathcal{P}_{\mathrm{X}}(x) = \mathcal{P}(\mathsf{E}_{X=x})$ is a function of x—the probability mass function. The right side of this definition is a function of x.

In the continuous case, no outcomes have X exactly equal to any particular chosen value. But we can define the cumulative event $\mathsf{E}_{X\leq x}$, and from that the probability density function:

$$\wp_{\mathrm{X}}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{P}(\mathsf{E}_{X\leq x}). \tag{5.20}$$

This definition makes it clear that $\wp_{\mathrm{X}}(x)$ is a function of x with dimensions inverse to those of x. (Integrating both sides and using the fundamental theorem of calculus shows that the cumulative distribution is the integral of the pdf.) For a joint distribution, we generalize to

$$\wp_{\mathrm{X,Y}}(x, y) = \frac{\partial^2}{\partial x \partial y}\mathcal{P}(\mathsf{E}_{X\leq x} \textbf{ and } \mathsf{E}_{Y\leq y}).$$

The formulas just outlined assume that a "probability measure" $\mathcal{P}$ has already been given that assigns a number to each of the events $\mathsf{E}_{X\leq x}$. It does not tell us how to *find* that measure in a situation of interest. For example, we may have a replicable system in which a single numerical quantity x is a complete description of what was measured; then Equation 5.1 effectively defines the probability. Or we may have a quantifiable degree of belief in the plausibility of various values of x (see Section 6.2.2).

Your Turn 5I

Rederive Equation 5.16 from Equation 5.20.

T_2 **Track 2**

5.2.4′ Interquartile range

The main text noted that the variance of a distribution is heavily influenced by outliers, and is not even defined for certain long-tail distributions, including Cauchy. The

text also pointed out that another measure of spread, the FWHM, is usable in such cases.

Another widely used, robust measure of the spread of a one-variable distribution is the **interquartile range** (IQR). Its definition is similar to that of the median:[27] We start at the lower extreme of the values obtained in a data sample and work our way upward. Instead of stopping when we have seen half of the data points, however, we stop after 1/4; this value is the lower end of the IQR. We then continue upward until we have seen 3/4 of the data points; that value is the upper end. The range between these two limits contains half of the data points and is called the interquartile range.

A more primitive notion of spread is simply the **range of a dataset**, that is, the difference between the highest and lowest values observed. Although any finite number of observations will yield a value for the range, as we take more and more data points the range is even more sensitive to outliers than the variance; even a Gaussian distribution will yield infinite range as the number of observations gets large.

T_2

Track 2

5.4′a Terminology

Power-law distributions are also called Zipf, zeta, or Pareto distributions in various contexts. Sometimes these terms are reserved for the situation in which $\wp$ equals $Ax^{-\alpha}$ exactly for x greater than some "cutoff" value (and it equals zero otherwise). In contrast, the Cauchy distribution is everywhere nonzero but deviates from strict power-law behavior at small $|x|$; nevertheless, we will still call it an example of a power-law distribution because of its asymptotic form at large x.

5.4′b The movements of stock prices

A stock market is a biological system with countless individual actors rapidly interacting with one another, each based on partial knowledge of the others' aggregate actions. Such systems can display interesting behavior.

It may seem hopeless to model such a complex system, and yet its very complexity may allow for a simplified picture. Each actor observes events external to the market (politics, natural disasters, and so on), adjusts her optimism accordingly, and also observes other actors' responses. Predictable events have little effect on markets, because investors have already predicted them and factored them into market prices before they occur. It's the *unexpected* events that trigger big overall changes in the market. Thus, we may suspect that changes in a stock market index are, at least approximately, random in the sense of "no discernible, relevant structure" (Idea 3.2, page 39).

More precisely, let's explore the hypothesis that

> ***H0:*** *Successive fractional changes in a stock market index at times separated by some interval Δt are independent, identically distributed draws from some unknown, but fixed, distribution.* (5.21)

[27] See Problem 5.2.

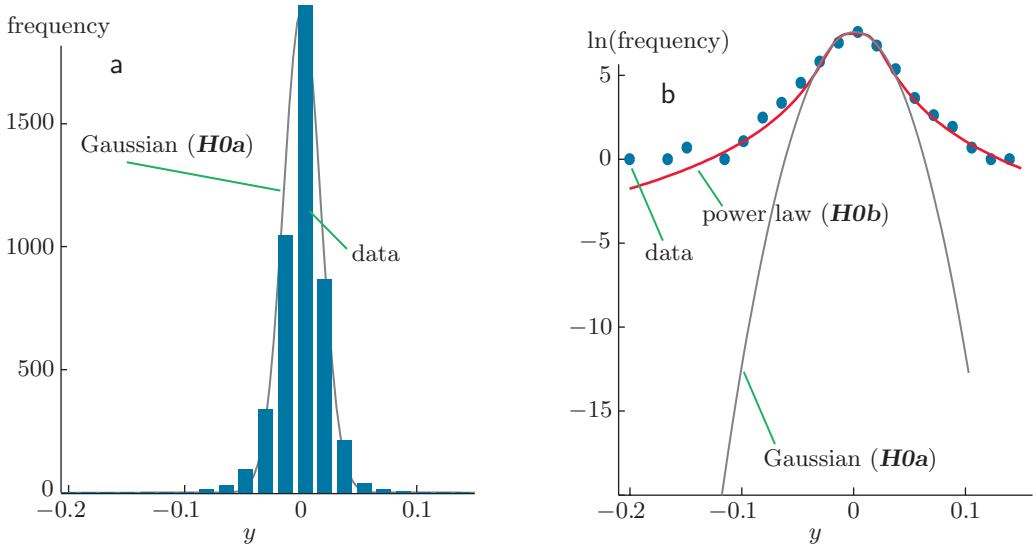


Figure 5.7 [Experimental data with fits.] **Another example of a long-tail distribution.** (a) *Bars:* Histogram of the quantity $y = \ln(x(t + \Delta t)/x(t))$, where x is the adjusted Dow Jones Industrial Average over the period 1934–2012 and $\Delta t = 1$ week. *Curve:* Gaussian distribution with center and variance adjusted to match the main part of the peak in the histogram. (b) *Dots and gray curve:* The same information as in (a), in a semilog plot. Some extreme events that were barely visible in (a) are now clearly visible. *Red curve:* Cauchy-like distribution with exponent 4. [Data from Dataset 7.]

This is not a book about finance; this hypothesis is not accurate enough to make you rich.[28] Nevertheless, the tools developed earlier in this chapter do allow us to uncover an important aspect of financial reality that was underappreciated for many years. To do this, we now address the specific question: *Under hypothesis **H0**, what distribution best reflects the available data?*

First note how the hypothesis has been phrased. Suppose that in one week a stock average changes from x to x', and that at the start of that week you owned \$100 of a fund tracking that average. If you did nothing, then at the end of that week the value of your investment would have changed to $\$100 \times (x'/x)$. A convenient way to express this behavior without reference to your personal circumstance is to say that the logarithm of your stake shifted by $y = \ln x' - \ln x$. Accordingly, the bars in Figure 5.7a show the distribution of historical y values.

It may seem natural to propose the more specific hypothesis ***H0a*** that the distribution of y is Gaussian. After all, a stock average is the weighted sum of a very large number of individual stock prices, each in turn determined by a still larger number of transactions in the period Δt, so "surely the central limit theorem implies that its behavior is Gaussian." The figure does show a pretty good-looking fit. Still, it may be worth examining other hypotheses. For example, the Gaussian distribution predicts negligible probability of $y = -0.2$, and yet the data show that such an event did occur. It matters: In an event of that magnitude, investors lost $1 - e^{-0.2} = 18\%$ of their investment value in a single week.

Moreover, each investor is *not at all* independent of the others' behavior. In fact, large market motions are in some ways like avalanches or earthquakes (Figure 5.6e): Strains

[28] For example, there are significant time correlations in real data. We will minimize this effect by sampling the data at the fairly long interval of $\Delta t = 1$ week.

build up gradually, then release suddenly as many investors simultaneously change their level of confidence. Section 5.4 pointed out that such collective-dynamics systems can have power-law distributions. Indeed, Figure 5.7b shows that a distribution of the form $\wp(y) = A/\left[1 + \left((y - \mu_y)/\eta\right)^4\right]$ (hypothesis ***H0b***) captures the extreme behavior of the data much better than a Gaussian.

To decide which hypothesis is *objectively* more successful, you'll evaluate the corresponding likelihoods in Problem 6.10. At the same time, you'll obtain the objectively best-fitting values of the parameters involved in each hypothesis family. For more details, see Mantegna and Stanley (2000).

PROBLEMS

5.1 Data lumping

Suppose that the body weight x of individuals in a population is known to have a pdf that is **bimodal** (two bumps; see Figure 5.8). Nora measured x on a large population. To save time on data collection, she took individuals in groups of 10, found the sample mean value of x for each group, and recorded only those numbers in her lab notebook. When she later made a histogram of those values, she was surprised to find that they didn't have the same distribution as the known distribution of x. Explain qualitatively what distribution they did have, and why.

5.2 Median mortality

The **median** of a random variable x can be defined as the value $x_{1/2}$ for which the probability $\mathcal{P}(x < x_{1/2})$ equals $\mathcal{P}(x > x_{1/2})$. That is, a draw of x is equally probable to exceed the median as it is to fall below it.

In his book *Full House*, naturalist Stephen Jay Gould describes being diagnosed with a form of cancer for which the median mortality, at that time, was eight months—in other words, of all people with this diagnosis, half would die in that time. Several years later, Gould noticed that he was alive and well. Can we conclude that the diagnosis was wrong? Sketch a possible probability distribution to illustrate your answer.

5.3 Cauchy-like distribution

Consider the family of functions

$$A\left(1+\left|\frac{x-\mu_x}{\eta}\right|^{\nu}\right)^{-1}.$$

Here μ_x, A, η, and ν are some constants; A, η, and ν are positive. For each choice of these parameters, we get a function of x. Could any of these functions be a legitimate probability density function for x in the range from $-\infty$ to ∞? If so, which ones?

5.4 Simulation via transformation

Section 5.2.6 explained how to take Uniformly distributed random numbers supplied by a computer and convert them into a modified series with a more interesting distribution.

a. As an example of this procedure, generate 10 000 Uniformly distributed real numbers x between 0 and 1, find their reciprocals $1/x$, and make a histogram of the results. (You'll

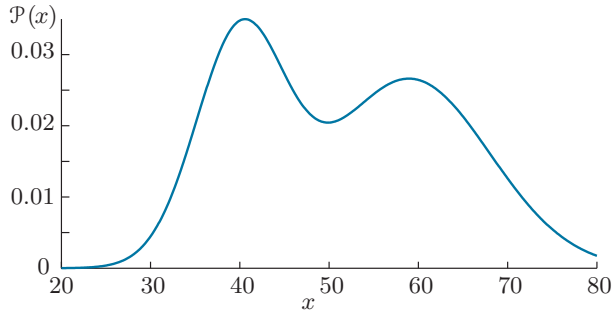


Figure 5.8 [Mathematical function.] **See Problem 5.1.**

need to make some decisions about what range of values to histogram and how many bins to use in that range.)

b. The character of this distribution is revealed most clearly if the counts in each bin (observed frequencies) are presented as a log-log plot, so make such a graph.

c. Draw a conclusion about this distribution by inspecting your graph. Explain mathematically why you got this result.

d. Comment on the high-x (lower right) end of your graph. Does it look messy? What happens if you replace 10 000 by, say, 50 000 samples?

5.5 Variance of Cauchy distribution

First work Your Turn 5G (page 106). Then consider the Cauchy distribution, Equation 5.9, with $\mu_x = 0$ and $\eta = 1$. Generate a set of 4 independent simulated draws from this distribution, and compute the sample mean of x^2. Now repeat for sets of 8, 16, . . . 1024, . . . draws and comment in the light of the Example on page 103.

5.6 Binomial to Gaussian (numerical)

Write a computer program to generate graphs like Figure 5.3b. That is, plot the Binomial distribution, properly rescaled to display the collapse onto the Gaussian distribution in an appropriate limit.

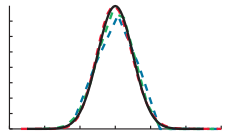


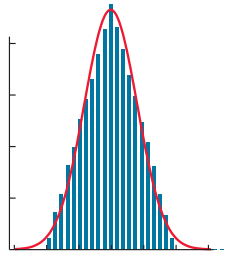
Figure 5.3b (page 107)

5.7 Central limit

a. Write a computer program to generate Figures 5.4a,b. That is, use your computer math package's Uniform random generator to simulate the two distributions shown, and histogram the result. Superimpose on your histograms the continuous probability density function that you expect in this limit.

b. Repeat for sums of 10 Uniform random numbers. Also make a semilog plot of your histogram and the corresponding Gaussian in order to check the agreement more closely in the tail regions. [*Hint:* You may need to use more than 50 000 samples in order to see the behavior out in the tails.]

c. Try the problem with a more exotic initial distribution: Let x take only the discrete values 1, 2, 3, 4 with probabilities 1/3, 2/9, 1/9, and 1/3, respectively, and repeat (a) and (b). This is a bimodal distribution, even more unlike the Gaussian than the one you tried in (a) and (b).



Figures 5.4a (page 108)

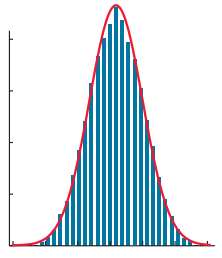


Figure 5.4b (page 108)

5.8 Transformation of multivariate distribution

Consider a joint probability density function for two random variables given by

$$\wp_{x,y}(x, y) = \wp_{\text{gauss},x}(x; 0, \sigma)\wp_{\text{gauss},y}(y; 0, \sigma).$$

Thus, x and y are independent, Gaussian-distributed variables.

a. Let $r = \sqrt{x^2 + y^2}$ and $\theta = \tan^{-1}(y/x)$ be the corresponding polar coordinates, and find the joint pdf of r and θ.

b. Let $u = r^2$, and find the joint pdf of u and θ.

c. Connect your result in (b) to what you found by simulation in Problem 3.4.

d. $\boxed{T_2}$ Generalize your result in (a) for the general case, where (x, y) have an arbitrary joint pdf and (u, v) are an arbitrary transformation of (x, y).

5.9 Power-law distributions

Suppose that some random system gives a continuous numerical quantity x in the range $1 < x < \infty$, with probability density function $\wp(x) = Ax^{-\alpha}$. Here A and α are positive constants.

a. The constant A is determined once α is specified. Find this relation.

b. Find the expectation and variance of x, and comment.

5.10 Tail probabilities

In this problem, you will explore the probability of obtaining extreme ("tail") values from a Gaussian versus a power-law distribution. "Typical" draws from a Gaussian distribution produce values within about one standard deviation of the expectation, whereas power-law distributions are more likely to generate large deviations. Your job is to make this intuition more precise. First work Problem 5.3 if you haven't already done so.

a. Make a graph of the Cauchy distribution with $\mu_x = 0$ and $\eta = 1$. Its variance is infinite, but we can still quantify the width of its central peak by the full width at half maximum (FWHM), which is defined as twice the value of x at which $\wp$ equals $\frac{1}{2}\wp(0)$. Find this value.

b. Calculate the FWHM for a Gaussian distribution with standard deviation σ. What value of σ gives the same FWHM as the Cauchy distribution in (a)? Add a graph of the Gaussian with this σ and expectation equal to zero to your graph in (a), and comment.

c. For the Cauchy distribution, calculate $\mathcal{P}(|x| > \text{FWHM}/2)$.

d. Repeat (c) for the Gaussian distribution you found in (b). You will need to do this calculation numerically, either by integrating the Gaussian distribution or by computing the "error function."

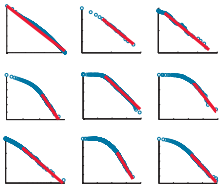


Figure 5.6 (page 111)

e. Repeat (c) and (d) for more extreme events, with $|x| > \frac{3}{2}\text{FWHM}$.

f. T_2 Repeat (a)–(e) but use the interquartile range instead of FWHM (see Section 5.2.4′).

5.11 Gaussian versus power law

To understand the graphs in Figure 5.6 better, consider the following three pdfs:

a. $\wp_1(x)$: For $x > 0$, this pdf is like the Gaussian distribution centered on 0 with $\sigma = 1$, but for $x < 0$ it's zero.

b. $\wp_2(x)$: For $x > 0$, this pdf is like the Cauchy distribution centered on zero with $\eta = 1$, but for $x < 0$ it's zero.

c. $\wp_3(x)$: This pdf equals $(0.2)x^{-2}$ for $x > 0.2$ and is zero elsewhere.

For each of these distributions, find the complementary cumulative distribution, display them all on a single set of log-log axes, and compare with Figure 5.6.

5.12 T_2 Convolution of Gaussians

Section 4.3.5 (page 79) described an unusual property of the Poisson distributions: The convolution of any two is once again Poisson. In this problem, you will establish a related result, in the domain of continuous distributions.

Consider a system with two independent random variables, x and y. Each is drawn from a Gaussian distribution, with expectations μ_x and μ_y, respectively, and variances both equal to σ^2. The new variable $z = x + y$ will have expectation $\mu_x + \mu_y$ and variance $2\sigma^2$.

a. Compute the convolution integral of the two distributions to show that the pdf $\wp_z$ is in fact precisely the Gaussian with the stated properties. [*Note:* This result is implicit in

the much more difficult central limit theorem, which roughly states that a distribution becomes "more Gaussian" when convolved with itself.]

b. Try the case where σ_x and σ_y are not equal.

5.13 T2 Convolution of Cauchy

a. Consider the Cauchy distribution with $\eta = 1$ and $\mu = 0$. Find the convolution of this distribution with itself.

b. What is the qualitative form of the convolution of this distribution with itself 2^p times? Comment in the light of the central limit theorem.

5.14 T2 Binomial to Gaussian (analytic)

This problem pursues the idea of Section 5.3.1, that the Gaussian distribution is a particular limit of the Binomial distribution.[29] You'll need a mathematical result known as **Stirling's formula**, which states that, for large M,

$$\ln(M!) \xrightarrow[M\to\infty]{} (M + \tfrac{1}{2}) \ln M - M + \frac{1}{2} \ln(2\pi) + \cdots . \tag{5.22}$$

The dots represent a correction that gets small as M gets large.

a. Instead of a formal proof of Equation 5.22, just try it out: Graph each side of Equation 5.22 for $M = 1$ to 30 on a single set of axes. Also graph the difference of the two sides, to compare them.

b. Following the main text, let $\mu_\ell = M\xi$ and $s = \sqrt{M\xi(1-\xi)}$, and define $y = (\ell - \mu_\ell)/s$. Then $\langle y \rangle = 0$ and $\operatorname{var} y = 1$, and the successive values of y differ by $\Delta y = 1/s$. Next consider the function

$$F(y; M, \xi) = (1/\Delta y)\mathcal{P}_{\text{binom}}(\ell; M, \xi), \text{ where } \ell = (sy + \mu_\ell).$$

Show that, in the limit where $M \to \infty$ holding fixed y and ξ, F defines a pdf. [*Hint:* s and ℓ also tend to infinity in this limit.]

c. Use Stirling's formula to evaluate the limit. Comment on how your answer depends on ξ.

5.15 T2 Wild ride

Obtain Dataset 7. The dataset includes an array representing many observations of a quantity x. Let $u_n = x_{n+1}/x_n$ be the fractional change between successive observations. It fluctuates; we would like to learn something about its probability density function. In this problem, neglect the possibility of correlations between different observations of u (the "random-walk" assumption).

a. Let $y = \ln u$; compute it from the data, and put it in an array. Plot a histogram of the distribution of y.

b. Find a Gaussian distribution that looks like the one in (a). That is, consider functions of the form

$$N_{\text{tot}} \frac{\Delta y}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(y-\mu_y)^2/(2\sigma^2)},$$

[29] The limit studied is different from the one in Section 4.3.2 because we hold ξ, not $M\xi$, fixed as $M \to \infty$.

for some numbers μ_y and σ, where Δy is the range corresponding to a bar on your histogram. Explain why we need the factors $N_{\text{tot}}\Delta y$ in the above formula. Overlay a graph of this function on your histogram, and repeat with different values of μ_y and σ until the two graphs seem to agree.

c. It can be revealing to present your result in a different way. Plot the logarithm of the bin counts, and overlay the logarithm of the above function, using the same values of σ and μ_y that you found in (a). Is your best fit really doing a good job? (Problem 6.10 will discuss how to make such statements precise; for now just make a qualitative observation.)

d. Now explore a different family of distributions, analogous to Equation 5.9 (page 101) but with exponents larger than 2 in the denominator. Repeat steps (a)–(c) using distributions from this family. Can you do a better job modeling the data this way?