1 Introduction

I know too well that these arguments from probabilities are impostors, and unless great caution is observed in the use of them they are apt to be deceptive – in geometry, and in other things too

(from Plato's Phaedo)

The purposes of this book are to familiarize you with a broad range of examples where randomness plays a key role, develop an intuition for it, and get to the level where you may read a recent research paper on the subject and be able to understand the terminology, the context, and the tools used. This is in a sense the "organizing principle" behind the various chapters: In all of them we are driven by applications where probability plays a fundamental role, and leads to exciting and often intriguing phenomena. There are many relations between the chapters, both in terms of the mathematical tools and in some cases in terms of the physical processes involved, but one chapter does not follow from the previous one by necessity or hinge on it – rather, the idea is to present a rich repertoire of problems involving randomness, giving the reader a good basis in a broad range of fields . . . and to have fun along the way.

Randomness leads to new phenomena. In a classic paper, Anderson (1972) coined the phrase "more is different". It is also true that "stochastic is different" . . . The book will give you some tools to understand phenomena associated with disordered systems and stochastic processes. These will include percolation (relevant for polymers, gels, social networks, epidemic spreading); random matrix theory (relevant for understanding the statistics of nuclear and atomic levels, model certain properties of ecological systems and more); random walks and Langevin equations (pertinent to understanding numerous applications in physics, chemistry, cell biology as well as finance). The emphasis will be on understanding the phenomena and quantifying them. Note that while all of the applications considered here build on randomness in a fundamental way, the collection of topics covered is far from representative of the vast realm of applications that hinge on probability theory. (For instance, two important fields not touched on here are statistical inference and chemical kinetics).

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

What mathematical background is assumed? The book assumes a solid background in undergraduate-level mathematics: primarily calculus, linear algebra, and basic probability theory. For instance, when a PDE (partial differential equation) is derived, we will not dwell too much on its solution if standard techniques are utilized, as will be the case when using standard results of linear algebra (e.g., that a Hermitian matrix admits unitary diagonalization). Similarly, if Lagrange multipliers are needed to perform a minimization, familiarity with them will be *assumed*. We will occasionally evaluate integrals using contour integration, though a motivated reader will be able to follow the vast majority of the book without a background in complex analysis. A summary/refresher of some of the techniques utilized is provided in Appendices A–E. A concise mathematical physics textbook that may come in handy for a reader who needs a further reminder of the techniques is by Mathews and Walker (1970). For readers who need to fill in a gap in their mathematical background (e.g., complex analysis), two excellent textbooks that cover the vast majority of the mathematical background assumed here (and much more) are by Hassani (2013) and Arfken, Weber, and Harris (2012). Further references for particular subject matter (probability theory, linear algebra, etc.) are provided in the Appendices.

To Prove or Not to Prove ... That Is the Question It is also important to emphasize what this book is **not** about: The derivations we will present will *not* be mathematically rigorous. What a physicist considers a proof is not what a mathematician would! In many cases, the physicist's proof has to be later "redone" by a mathematician, sometimes decades later. However, for various real-world applications the advantages of a rigorous proof might not be justifiable. Quoting Feynman, "If there is something very slightly wrong in our definition of the theories, then the full mathematical rigor may convert these errors into ridiculous conclusions." To paraphrase Feynman – in many cases, there is no need to solve a model exactly, since the connection between the model and the reality is only crude, and we made numerous (far more important) approximations in deriving the model equations (von Neumann put this more bluntly: "There's no sense in being precise when you don't even know what you're talking about"). Similar expectations will hold for the exercises at the end of each chapter, which also consist of numerical simulations in cases where analytic derivations are impossible or outside the scope of the book.

Furthermore, the notation and jargon we will follow will be those used by physicists in research papers – which is sometimes different from those used by mathematicians (for instance, we will refer to the object mathematicians call the "probability density function" as the "probability distribution", and use the physicists' notation for it – $p(x)$). We will often not specify the precise mathematical conditions under which the derivations can be made rigorous, and we will shamelessly use algebraic manipulations without justifying them – such as using Fubini's theorem without ensuring the function is absolutely integrable. All functions will be assumed to be differentiable as many times as necessary. We will also be using Dirac's δ-function in the derivations,

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

and creatures such as the Fourier transform of $e^{i\omega t}$. For a physicist, these objects should be interpreted under the appropriate regularization – a δ-function should be thought of as having a finite width, but much smaller than any other relevant scale in the problem. (Physicists are relatively used to this sort of regularization – for instance, in computing Green functions using contour integration the contour often has to be shifted by an amount $\pm i\epsilon$ to make the results convergent). If the final result depends on this finite width – then the treatment using δ-function is inadequate and should be revisited. But as long as the final results are plausible (e.g., in some cases we can compare with numerics) we will not re-derive them in a rigorous fashion. The advantage of this non-rigorous approach is that it seems to be the one more relevant to applications, which are the focus of this book. Experiments and real-life phenomena do not conform to the mathematical idealization we make anyhow, and von Neuman's quote comes to mind again. In other words, the more important thing for explaining physical reality is to have a good model rather than specify the conditions rigorously (a related quote is attributed to Kolmogorov: "Important is not what is rigorous but what is true"). That is not to take anything away from the beautiful work of mathematicians – it is just not the point of this book.

For some students, this non-rigorous approach could prove challenging. When the rigorously inclined student encounters a situation where they feel the formal manipulations are unjustified, it may prove useful for them to construct counter-examples, e.g., functions which do not obey the theorem, and then to consider the physical meaning of these "good" and "bad" functions – which class is relevant in which physical situations, and what we learn from the scenarios where the derivation fails. Our goal here is not to undermine the importance of rigorous mathematics, but to provide a non-rigorous introduction to the plethora of natural sciences phenomena and applications where stochasticity plays a central role.

How to read this book A few words on the different topics covered and their relations. Chapter 1 is introductory, and gives some elementary examples where basic probability theory leads to perhaps counter-intuitive results. One of the examples, Benford's law, touches on some of the topics of Chapter 6 (dealing with heavy-tailed distributions, falling off as a power-law). Chapter 2 presents random walks and diffusion, and provides the foundational basis for many of the other chapters. Chapter 3 directly builds on the simple random walks introduced in Chapter 2, and discusses the important concepts of Langevin and Fokker–Planck equations. The first part of the chapter "builds" the formalism (albeit in a non-technical and non-rigorous fashion), while the second part of the chapter deals with three applications of the ideas (cell size control – an application in biology, the Black–Scholes equation – one in economics, and finally a short application in hydrology). A reader may skip these applications without affecting the readability of the rest of the materials. Similarly, Chapter 4 (dealing with the "escape over a barrier" problem) can be viewed as a sophisticated application of the ideas of Chapter 3, with far-reaching implications. It certainly puts the materials of the previous chapters to good use, but again can be skipped without affecting the flow. Chapter 5

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

is of particular importance to those dealing with signals and noise, and builds on ideas introduced in earlier chapters (e.g., the Markov chains of Chapter 2) to analyze the power spectrum (i.e., noise characteristics) of several paradigmatic systems (including white noise, telegraph noise, and $1/f$ noise). Chapter 6 derives a plethora of basic results dealing with the central limit theorem, its limitations and generalizations, and the related problem of "extreme value distributions". It is more technical (and lengthier) than previous chapters. Chapter 7, dealing with anomalous diffusion, can be viewed as an advanced application of the materials of Chapter 6, "reaping the fruits" of the labor of the previous chapter. In a sense, it extends the results of the random walks of Chapter 2 to scenarios where some of the assumptions of Einstein's approach do not hold – and have been shown to be relevant to many systems in physics and biology (reminiscent of the quote, "everything not forbidden is compulsory" . . .) Chapter 8 deals with random matrices and some of their applications. It is the most technical chapter in this book, and is mostly independent from the chapter on percolation theory that follows. Moreover, a large fraction of Chapter 8 deals with a non-trivial derivation of the "circular law" associated with non-Hermitian matrices, and a reader can skip directly to Chapter 9 if they prefer. (Note that most of this lengthy derivation "unzips" the short statements made in the original paper, perhaps giving students a glimpse into the compact nature in which modern research papers are written!) The final chapter on percolation theory touches on fundamental concepts such as emergent behavior, the renormalization group, and critical phenomena. Throughout the chapters, numerical simulations in MATLAB are provided when relevant.* Often, results are easy to obtain numerically but challenging to derive analytically, highlighting the importance of the former as a supplement to analytic approaches. Finally, note that occasionally "boxes" are used where we emphasize an idea or concept by placing the passage between two solid lines.

Note that each chapter deals with a topic on which many books and many hundreds of papers have been written. This book merely opens a narrow window into this vast literature. The references throughout the book are also by no means comprehensive, and we apologize for not including numerous relevant references – this text is not intended to be a comprehensive guide to the literature! When possible, we refer to textbooks on the topic that provide a more in-depth discussion as well as a more extensive list of references.

A comment on the problems in this book (and their philosophy) The problems at the end of each chapter are a little different from those encountered in most textbooks. The phrasing is often laconic or even vague. Students might complain that "the problem is not hard – I just cannot figure out what it is!" This actually reflects the typical situation in many real-life problems, be it in academia or industry, where figuring out how to *set up* the problem is often far more challenging than solving the problem itself. The Google PageRank algorithm described in Chapter 2 is a nice example where simple, well-known linear algebra can be highly influential when used correctly in the appropriate context. The situation might be frustrating at times, when trying to

* The codes can be downloaded here: https://github.com/arielamir/ThinkingProbablistically

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

prove something without being given in advance the precise conditions for the results to hold – yet this mimics the situation encountered so often in research. Indeed, many of the original problems arose from the author's own research experience or from (often recent) research papers, and as such reflect "natural" problems rather than contrived exercises. In other cases, the problems supplement the materials of the main chapter and essentially "teach" a classic theorem (e.g., Pólya's theorem in Chapter 2) through hands-on experience and calculations (and with the proper guidance to make it manageable, albeit occasionally challenging). We made a conscious choice to make the problems less defined and avoid almost categorically problems of the form "Prove that X takes the form of Y under the assumptions Z." The philosophy behind this choice is to allow this book (and the problems) to serve as a bridge between introducing the concepts and doing research on related topics. The typical lack of such bridges is nicely articulated by Williams (2018), which was written by a graduate student based on his own first-hand experience in making the leap from undergraduate course work to graduate-level physics research. Trickier problems will be denoted by a * (hard) or ** (very hard), based on the previous experience of students tackling these problems.

1.1 Probabilistic Surprises

Randomness can lead to counter-intuitive effects and to novel phenomena

Research has shown that our intuition for probability is far from perfect. Problems associated with probability are often easy to formulate but hard to solve, as we shall see throughout this book. Many problems have an element of randomness to them (sometimes due to our imperfect knowledge of the system) and may be best examined via a probablisitic model.

As a "warmup," let us consider several elementary examples, where the naive expectation (of most people) fails, while a simple calculation gives a counterintuitive result. Some excellent examples that we will not consider – since they are perhaps too well known – include the Monty–Hall problem (http://en.wikipedia.org/wiki/MontyHallproblem), and the false-negative paradox (http://en.wikipedia.org/wiki/Falsepositiveparadox).

1.1.1 Example: Does the Fraction of a Rare Disease in a Population Remain Fixed Over Time? (Hardy–Weinberg Equilibrium)

Consider a rare disease associated with some recessive allele (i.e., the individual must have two copies of this gene, one from each parent, in order to be sick). The population consists of three genotypes: individuals with two copies of the dominant gene (who will not be sick), which we will denote by AA; those with one copy of the recessive gene, aA; and those with two copies of it aa, who will be sick. Each of the parents gives one of the alleles to the offspring, with equal probability. This is schematically described in the table below, describing the various possibilities for

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

the offspring's genotype (and their probabilities in brackets) given the mother's and father's genotypes.

mother \ father	**aa**	**AA**	**aA**
aa	$aa(1)$	$aA(1)$	$aA\left(\frac{1}{2}\right), aa\left(\frac{1}{2}\right)$
AA	$aA(1)$	$AA(1)$	$AA\left(\frac{1}{2}\right), aA\left(\frac{1}{2}\right)$
aA	$aa\left(\frac{1}{2}\right), aA\left(\frac{1}{2}\right)$	$AA\left(\frac{1}{2}\right), aA\left(\frac{1}{2}\right)$	$aa\left(\frac{1}{4}\right), AA\left(\frac{1}{4}\right), aA\left(\frac{1}{2}\right)$

We shall denote the relative abundance of genotypes AA, aA, and aa by p, $2q$, and r, respectively (thus, by definition $p + 2q + r = 1$). Surprisingly, at the beginning of the twentieth century, it was not clear what controls the relative abundance of the three types: *What are the possible stable states? What are the dynamics starting from a generic initial condition?*

On surprises If you haven't seen this problem before, you might have some prior intuition or guesses as to what the results might be. For instance, it might be reasonable to expect that if a disease corresponding to a recessive gene is initially very rare in the population, then over time it should go extinct. This is, in fact, *not* the case, as we shall shortly see. In that sense, you may call the results "surprising." But perhaps a reader with better intuition would have guessed the correct result a priori, and will not find the result surprising at all – in that sense, the notion of a "surprising result" in science is, in fact, a rather unscientific concept. In retrospect, mathematical results cannot really be surprising . . . Nevertheless, the scientific *process* itself is often driven by intuition and lacks the clarity of thought that is the luxury of hindsight, and for this reason scientists do often invoke the concept of a "surprising result." Moreover, this often reflects our expectations from prior null models that we are familiar with. For example, in Section 1.1.2 we will show a simple model suggesting an exponential distribution of the time intervals between subsequent buses reaching a station. Armed with this insight, we can say that the results described in Chapter 8, finding a distribution of time intervals between buses that is not only non-exponential but in fact non-monotonic, are surprising! But this again illustrates that our definition of surprising very much hinges on our prior knowledge, and perhaps a more (or less) mathematically sophisticated reader would not find the latter finding surprising. For a related paper, see also Amir, Lemeshko, and Tokieda (2016b).

Remarkably, it was not until 1908 that the mathematician G. H. Hardy sent a letter to the editor of *Science* magazine clearing up this issue (Hardy 1908). His letter became a cornerstone of genetics (known today as the Hardy–Weinberg equilibrium, a name also crediting the independent contributions of Wilhelm Weinberg). The model and calculations are extremely simple. Assuming a well-mixed population in its nth generation, let us compute the abundance of the three genotypes in the $n + 1$ generation, assuming for simplicity random mating between the three genotypes. Using the

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

table, it is straightforward to work out that the equations relating the fractions in one generation to the next are:

$$p_{n+1} = p^2 + 2pq + q^2 = (p+q)^2. \tag{1.1}$$

$$r_{n+1} = r^2 + 2qr + q^2 = (r+q)^2. \tag{1.2}$$

$$2q_{n+1} = 2q^2 + 2pq + 2qr + 2pr = 2(p+q)(r+q). \tag{1.3}$$

Note that we dropped the $_n$ subscript on the RHS (the abbreviation we will use for "right-hand side" through the text) to make the notation less cumbersome. As a sanity check, you can check that these sum up to $(p+2q+r)^2 = 1$.

If we reach a stationary ("equilibrium") state, then $p_{n+1} = p_n$, etc. This implies that

$$pr = q^2. \tag{1.4}$$

Hence q is the geometric mean of p and r at equilibrium. Is this a sufficient condition for equilibrium? The answer is yes, since the first equation becomes

$$p = p^2 + 2pq + pr = p(p+2q+r) = p \tag{1.5}$$

(and the second equation has the same structure – can you see why there is no need to check the third?).

Finally, how long would it take us to reach this state starting from general initial conditions p_1, q_1, and r_1? Note that $q_{n+1} = (p+q)(r+q)$, hence:

$$q_2^2 = (p_1+q_1)^2(r_1+q_1)^2 = p_2r_2.$$

So equilibrium is precisely established after a single generation! Importantly, in sharp contrast to the biologists' prior belief, it can have an arbitrarily small – but stable – value of r. Things are different, in fact, in the subtle variant of this problem studied in Problem 1.2.

1.1.2 Example: Bus Timings, Random vs. Ordered

In bus station A, buses are regularly sent off every 6 minutes. In station B, on the other hand, the manager throws a die every minute and sends off a bus if they get a "6."

Q: How many buses leave per day? What is the average time between buses?

It is clear that on average $60 \cdot 24/6 = 240$ buses will be sent out, in both cases. Hence the average time between buses is 6 minutes in both cases (in case A, the time between buses is, of course, always 6 minutes).

Q: If a person arrives at a random time to the bus station, how many minutes do they have to wait on average?

For simplicity, let us assume that the person arrived just after a potential bus arrival event.

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

What should we expect? In case A, it is equally probable for the waiting time to be $1, 2 \ldots, 6$ minutes, hence the average waiting time is 3.5 minutes. Try to think about case B. Given that the total number of buses per day and the average time between buses is identical, you might expect that the average waiting time would be identical too. This is not the case, as we shall shortly show: In case B, the average time between buses is also 6 minutes, but, perhaps counterintuitively, this is also the average waiting time!

To see this, let us assume that we got to the station at a random time. The probability to wait a minute until the next one is $1/6$. The probability for a 2 minute wait is $\frac{5}{6}\frac{1}{6}$, and more generally the probability to wait n minutes, p_n, is

$$p_n = (1-p)^{n-1} p \tag{1.6}$$

(with $p = 1/6$).

Therefore, the average waiting time is

$$\langle T \rangle = \sum p_n n = \sum_{n=1}^{\infty} n(1-p)^{n-1} p. \tag{1.7}$$

Without the n in front, this would be a geometric series. To deal with it, define $q \equiv 1 - p$, and note that

$$\sum_{n=0}^{\infty} q^n = 1/(1-q). \tag{1.8}$$

Taking the derivative with respect to q gives us the following relation:

$$\sum_{n=0}^{\infty} n q^{n-1} = 1/(1-q)^2. \tag{1.9}$$

Therefore, our sum in Eq. (1.7) equals

$$\langle T \rangle = p/(1-q)^2 = (1/6)/(1/6)^2 = 6. \tag{1.10}$$

Looking back, this makes perfect sense, since the fact that a bus just left does not “help” us regarding the next one – the process has no memory. Interestingly, this example is relevant for the physics of a (classical) model of electron transport, known as the Drude model – where our calculations imply that an additional factor of “2” should *not* be present in the final result.

What about the distribution of time gaps? It is given by Eq. (1.6), and is therefore *exponential*. This process is a simple example of a random process, and in the continuum limit where the time interval is vanishingly small this is known as a Poisson process (see Appendix A for the related *Poisson distribution*, describing the probability distribution of the *number* of events occurring within a fixed time interval).

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

A note on terminology Throughout this book, we will follow the physicists' terminology of referring to a probability density function (pdf) as a "probability distribution," and referring to a "cumulative distribution" for the cumulative distribution function (cdf). Moreover, for a real random variable X we will denote the probability distribution by $p(x)$, rather than the notation f_X often used in mathematics. **Further notational details are provided in Appendix F.**

What about real buses? An online blog analyzed the transportation system in London and showed that it is Poissonian (i.e., corresponds to the aforementioned random bus scheduling) (http://jasmcole.com/2015/03/02/two-come-along-at-once/). This implies that the system is not optimal (since we can get buses coming in "bunches," as well as very long waits). On the other hand, later in the book (Chapter 8) we will see a case where buses were not Poissonian but also not uniform – the distribution was very different from exponential (the Poisson case) but was not narrowly peaked (the uniform case). Interestingly, it vanished at zero separation – buses "repelled" each other. It was found to be well described by the results of random matrix theory, which we shall cover in Chapter 8.

1.1.3 Bertrand's Paradox: The Importance of Specifying the Ensemble

Consider an equilateral triangle inscribed in a circle. Suppose a chord of the circle is chosen at random. What is the probability that the chord is longer than a side of the triangle?

We shall now show three "reasonable" methods for choosing a random chord, which will give us 1/2, 1/3, and 1/4 as an answer, respectively. This is known as "Bertrand's paradox" (Bertrand 1907).

Method 1: Choosing the endpoints. Let us choose the two endpoints of the chord at random. The chord is longer than the side of the triangle in 1/3 of the cases – as is illustrated in Fig. 1.1

Method 2: Choosing the midpoint. What about if we choose a point randomly and uniformly in the circle, and define it to be the middle of the chord? From the

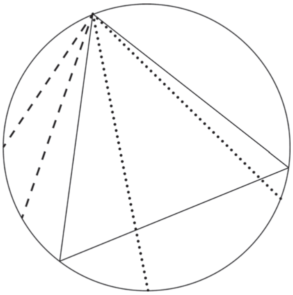

Figure 1.1 Method 1: Choosing the endpoints of the chord.

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

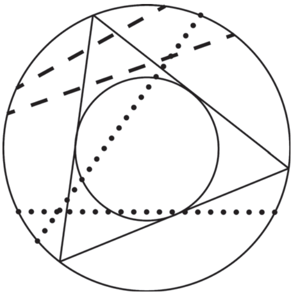

Figure 1.2 Method 2: Choosing the chord midpoint randomly and uniformly in the circle.

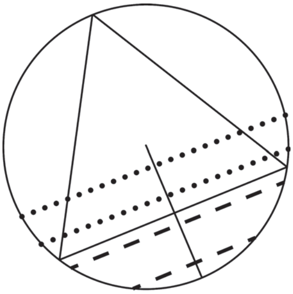

Figure 1.3 Method 3: Defining the chord midpoint by choosing a point along the radius.

construction of Fig. 1.2, we see that when the point falls within the inner circle the chord will be long enough. Its radius is $R\sin(30) = R/2$, hence its area is $1/4$ times that of the outer circle – therefore, the probability will be $1/4$.

Method 3: Choosing a point along the radius to define the midpoint. If we choose the chord midpoint along the radius of the circle with uniform probability, the chord will be long enough when the chosen point is sufficiently close to the center – it is easy to see that the triangle bisects the radius, so in this case the probability will be $1/2$ (see Fig. 1.3).

Importantly, there is no right or wrong answer – but the point is that one has to describe the way through which the "random" choice is made to fully describe the problem.

1.1.4 Benford's Law: The First Digit Distribution

Another related example where the lack of specification of the random ensemble leads to rather counterintuitive results is associated with Benford's law (named after Frank Benford, yet discovered by Simon Newcomb a few decades beforehand!). It is an empirical observation that for many "natural" datasets the distribution of the first digit is very far from uniform, see Fig. 1.4. The formula that the data is compared with is one where the relative abundance of the digit d is proportional to

$$p_d \propto \log(1 + 1/d) \tag{1.11}$$

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

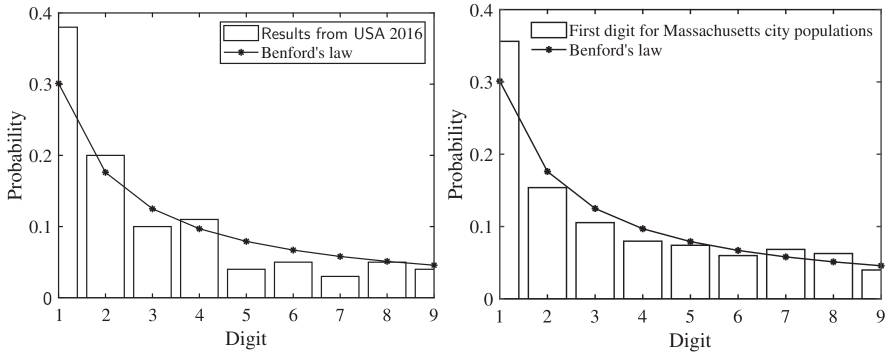

Figure 1.4 (left) Example of a dataset approximately following Benford's law, obtained by using readily available data for the vote total of the candidates in the 2016 elections in the USA across the different states (data from Wikipedia). You can easily test other datasets yourself. A similar analysis was used to suggest fraud in the 2009 Iranian elections (Battersby 2009). (right) Similar analysis of city population size in Massachusetts, based on the 2010 census data (www.togetherweteach.com/TWTIC/uscityinfo/21ma/mapopr/21mapr.htm).

(this implies that 1 occurs in about 30 % of cases and 9 in less than 5!) Although here it makes no difference, log refers to the natural logarithm throughout the text, unless otherwise specified.

Clearly, this is not always true, e.g., MATLAB's random number generator closely follows a uniform distribution, and hence the distribution of the first digit will be uniform. But it turns out to be closely followed for, e.g., tax returns, city populations, election results, physics constants, etc. What do these have in common? The random variable is broadly distributed, i.e., the distribution spans many decades, and it is far from uniform. In fact, to get Eq. (1.11), we need to assume that the *logarithm* of the distribution is uniformly distributed over a large number of decades, as we shall now show. If x is such a random variable, with $y = \log(x)$, then for values of y within the support of the uniform distribution we have

$$p(x)dx = g(y)dy = Cdy, \tag{1.12}$$

where $p(x)$ and $g(y)$ are the probability distributions, and $C = \frac{1}{y_{max} - y_{min}}$. Hence:

$$p(x) = C|dy/dx| = C/x \tag{1.13}$$

(see Appendix A for a reminder on working with such a change of variables). Therefore, $p(x)$ rapidly increases at low values of x (until dropping to zero at values of x below the lower cutoff – set by the cutoff of the assumed uniform distribution of y).

The relative abundance of the digit 1 is therefore

$$p_1 = \int_1^2 p(x)dx + \int_{10}^{20} p(x)dx \ldots \approx C \cdot D \log(2/1), \tag{1.14}$$

where D is the number of decades spanned by the distribution $p(x)$. Similarly

$$p_d \propto \int_d^{d+1} \frac{1}{x} dx = \log(1 + 1/d), \tag{1.15}$$

and Benford's law follows.

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

The importance of specifying the ensemble Although the problems are very different in nature, there is a deep analogy between Bertrand's paradox and Benford's law in the following sense: In both cases the "surprising result" comes from a lack of definition of the random ensemble involved. In Bertrand's paradox case, it is due to our loose phrasing of "random." In the case of Benford's law, it is manifested in our misconception that given that we are looking at a random variable, the distribution of the first digit should be uniform – this would indeed be true if the random variable is drawn from a broad, *uniform* distribution, but such distributions typically do not correspond to naturally occurring datasets.

It is easy to show that Benford's law arises if we assume that the distribution of the first digit is *scale invariant* (e.g., the tax returns can be made in dollars or euros). But why do such broad distributions arise in nature so often? One argument that can be made to rationalize Benford's law relies on *multiplicative processes*, an idea that (potentially) traces back to Shockley. He noticed that the productivity of physicists at Bell labs – quantified in terms of their publication number – follows a log-normal distribution, and wanted to rationalize this observation (Shockley 1957). We repeat the argument here, albeit for the example of the population of a city, x.

This depends on a large number N of "independent" variables: the availability of water, the weather, the quality of the soil, etc. If we assume that:

$$x = \prod_{i=1}^{N} x_1 \cdot x_2 \dots x_N, \tag{1.16}$$

with x_i some random, independent variables (e.g., drawn from a Gaussian distribution), then the distribution of the *logarithm* of x can be approximated by a Gaussian (by the central limit theorem, see Appendix A), hence $p(x)$ will be a log-normal distribution – which is very similar to the uniform distribution we assumed above, and Benford's law will approximately follow.

For another example where a similar argument is invoked to explain the observation of a log-normal distribution of file sizes, see Downey (2001). Another example of a variant of this argument relates to the logarithmic, slow relaxations observed in nature: see Amir, Oreg, and Imry (2012). We will revisit this example in Chapter 6 in more detail. Finally, the log-normal distribution and the multiplicative mechanism outlined above also pop up in the "numerology" context of Amir, Lemeshko, and Tokieda (2016b).

1.2 Summary

In this chapter we introduced a somewhat eclectic set of problems where probability leads to (perhaps) surprising results. Hardy–Weinberg involved genetics and population dynamics, where we derived a necessary and sufficient condition for the

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

relation of the three possible genotypes. We discussed the important example of a Poisson process, motivated by considering a simple model for the statistics of buses arriving at a station. Finally, we exemplified the huge importance of specifying the random ensembles we are dealing with, first by following Bertrand and considering a geometric “paradox” (resulting from our incomplete definition of the problem), and next by discussing (and attempting to rationalize) the empirical observation of a rather universal (but non-uniform) distribution of the first digit associated with naturally occurring datasets.

For further reading See Mlodinow (2009) for an elementary but amusing book on surprises associated with probability and common pitfalls, with interesting historical anecdotes.

1.3 Exercises

1.1 Probability Warm-up

In a certain village, each family has children until their first daughter is born, after which they stop. What is the ratio of the expected number of boys to girls in the village? Assume a girl or boy is born with probability $1/2$.

1.2 Hardy–Weinberg Revisited

Consider a rare, *lethal* genetic disease, such that those who carry two copies of the recessive gene (aa) will not live to reproduce.

(a) Derive the equations connecting the abundances of the AA, Aa, and aa genotypes (p, $2q$, and r) between subsequent generations.

(b) What happens at long times?

1.3 Benford’s Law and Scale Invariance

Prove that Benford’s law is a consequence of scale invariance: If the first digit distribution does not depend on units (e.g., measuring income in dollars vs euros), Benford’s law follows.

1.4 Benford’s Law for Second Digit

(a) Using the same assumptions behind the derivation of Benford’s law, derive the probability distribution of the first two digits of each number.

(b) What is the probability of finding 3 as the second digit of a number following Benford’s law?

1.5 Drude Model of Electrical Conduction

The microscopic picture of the Drude model is that, while moving through a metal, electrons can randomly collide with ions. Interactions between electrons are neglected, and electron–ion interactions are considered short range. The probability of a single electron collision in any infinitesimal interval of time of duration dt is dt/τ (and the probability of no collision is $1 - dt/\tau$).

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press

(a) Consider a given electron in the metal. Determine the probability of the electron not colliding within the time interval $[0, t]$.

(b) Assume that a given electron scatters at $t = 0$ and let T be the time of the following scattering event. Find the probability of the event that $t < T < t + dt$. Also calculate the expected time $\langle T \rangle$ between the two collisions.

(c) Let $t = 0$ be an arbitrary observation time. Let T_n be the time until the next collision after $t = 0$ and T_l be the time since the last collision before $t = 0$; Consider the random variable $T = T_l + T_n$. Determine the distributions of T_l and T_n, and from them deduce $\langle T \rangle$. Does your answer agree with the result in part (b)? If not, explain the discrepancy between the two answers.

1.6 Mutating Genome*

Consider an organism with a genome in which mutations happen as a Poisson process with rate U. Assume that all mutations are neutral (i.e., they do not affect the rate of reproduction). Assume the genome is large enough that the mutations always happen at different loci (this is known as the infinite sites model) and are irreversible. We start at $t = 0$ when there are no mutations.

(a) What is the probability that the genome does not obtain any new mutations within the time interval $[0, t)$?

(b) What is the expected number of mutations for time T?

(c) Consider a population following the Wright–Fisher model: At each generation, each of N individuals reproduces, but we keep the population size fixed by randomly sampling N of the $2N$ newborns. Find the probability of two individuals having their "first" (latest chronologically) common ancestor t generations ago, $P(t)$. Hint: Go backwards in time with discrete time steps. What is the continuum limit of this probability? (i.e., the result for a large population size).

(d) Let us add mutations to the Wright–Fisher model. Assume we sample two individuals that follow two different lineages for precisely t generations (i.e., their first common ancestor occured t generations ago). What is $P(\pi|t)$, the probability of π mutations arising during the t generations?

(e) What is $P(\pi)$, the probability of two individuals being separated by π mutations after they were born from the same parent? What is the expected value of π? (You may work in the continuum limit as in (c), corresponding to a large population size $N \gg 1$).

Problem credit: Jiseon Min.

https://doi.org/10.1017/9781108855259.002 Published online by Cambridge University Press