Contents

1

Introduction

This is a book about the Monte Carlo method. The core idea of Monte Carlo is to learn about a system by simulating it with random sampling. That approach is powerful, flexible and very direct. It is often the simplest way to solve a problem, and sometimes the only feasible way.

The Monte Carlo method is used in almost every quantitative subject of study: physical sciences, engineering, statistics, finance, and computing, including machine learning and graphics. Monte Carlo is even applied in some areas, like music theory, that are not always thought of as quantitative.

The Monte Carlo method is both interesting and useful. In this book we will look at the ideas behind Monte Carlo sampling and relate them to each other. We will look at the mathematical properties of the methods to understand when and why they work. We will also look at some important practical details that we need to know in order to get reliable answers to serious problems. Often the best way to see the ideas and practical details is through an example, and so worked examples are included throughout.

1.1 Example: traffic modeling

Monte Carlo methods have proven useful in studying how vehicle traffic behaves. Perhaps the most interesting aspect of traffic is the occurrence of traffic jams. At places where the number of traffic lanes is reduced, cars slow down and form a blockage. Similarly, accidents or poor visibility or the occasional slow vehicle can bring about a traffic jam.

Sometimes, however, a traffic jam has no apparent cause. It just spontaneously appears in flowing traffic, and moves slowly backwards against the traffic. It can last a long time and then simply disappear.

The phenomenon can be illustrated with Monte Carlo methods. A very simple Monte Carlo simulation that captures some of the important properties of real traffic is the Nagel-Schreckenberg model. In this model the roadway is divided up into M distinct zones, each of which can hold one vehicle. There are N vehicles in the road. Time moves forward in discrete steps. A vehicle with velocity v moves ahead by v zones in the roadway at the next time step. There is a maximum speed $v_{\max}$ which all vehicles obey. In the simplest case, the roadway is a circular loop.

The rules for ***Nagel-Schreckenberg traffic*** are as follows. At each stage of the simulation, every car goes through the following four steps. First, if its velocity is below $v_{\max}$, it increases its velocity by one unit. The drivers are eager to move ahead. Second, it checks the distance to the car in front of it. If that distance is d spaces and the car has velocity $v \geqslant d$ then it reduces its velocity to $d-1$ in order to avoid collision. Third, if the velocity is positive then with probability p it reduces velocity by 1 unit. This is the key step which models random driver behavior. At the fourth step, the car moves ahead by v units to complete the stage. These four steps take place in parallel for all N vehicles.

Let $x \in \{0, 1, \ldots, M-1\}$ be the position of a car, v its velocity, and d be the distance to the car ahead. The update for this car is:

$$\begin{aligned} v &\leftarrow \min(v+1, v_{\max}) \\ v &\leftarrow \min(v, d-1) \\ v &\leftarrow \max(0, v-1) \quad \text{with probability } p \\ x &\leftarrow x + v. \end{aligned} \tag{1.1}$$

At the last step, if $x + v \geqslant M$ then $x \leftarrow x + v - M$. Similarly, for the car with the largest x, the value of d is M plus the smallest x, minus the largest x.

Figure 1.1 shows a sample realization of this process. It had $N = 100$ vehicles in a circular track of $M = 1000$ spaces. The speed limit was $v_{\max} = 5$ and the probability of random slowing was $p = 1/3$. The initial conditions are described on page 10 of the chapter end notes. The figure clearly shows some traffic jams spontaneously appearing, then drifting backward, and then disappearing. We can also see some smaller congestions which move slowly forward over time. The white stripes are unusually wide gaps between cars. These gaps move at nearly the speed limit.

Obviously, real traffic is much more complicated than this model. The Nagel-Schreckenberg model has been extended to have multiple lanes. It and other models can be applied on arbitrarily large road networks for cars having different random sources, destinations and speeds. The much more realistic models can be used to make predictions of the effects of adding a lane, temporarily closing a city street, putting metering lights at a freeway on-ramp, and so on. Adding such realism requires considerable domain knowledge about traffic and lots of computer code and documentation. Monte Carlo methods are then a small but important ingredient in the solution.

The extreme simplicity of the Nagel-Schreckenberg model provides one advantage. There is no bottleneck, no accident, or other phenomenon to explain

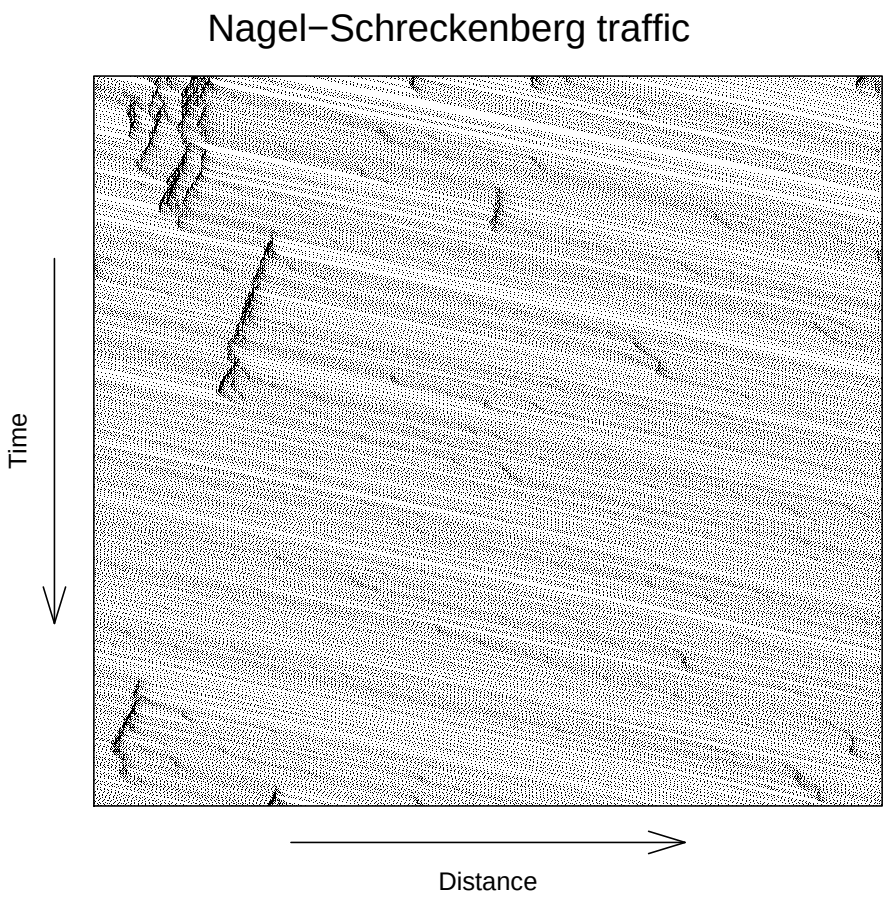

Figure 1.1: This figure illustrates a Monte Carlo simulation of the Nagel-Schreckenberg traffic model as described in the text. The simulation starts with 100 cars represented as black dots on the top row. The vehicles move from left to right, as time increases from top to bottom. Traffic jams emerge spontaneously. They show up as dark diagonal bands. The traffic jams move backward as the traffic moves forward.

the traffic jams. There is no particularly bad driver, because they all follow the same algorithm. Even without any of that complexity, a bit of random slowing is enough to cause stop-and-go traffic patterns to emerge and then dissipate.

The other advantage of very simple models is pedagogic. They bring the Monte Carlo issues to the forefront. A full description of the non-Monte Carlo issues in a realistic simulation could well be larger than this whole book.

In the traffic example, the emphasis is on making a picture to get qualitative

insight. Such visualization is a very common use of Monte Carlo methods. Sometimes the picture is the goal in itself. For example, Monte Carlo methods are widely used in the making of movies, and Oscars have even been awarded for progress in Monte Carlo methods.

Usually when we see a feature in a picture we want a quantitative measure of it. It is more typical for Monte Carlo investigations to be aimed at getting a number, rather than making a picture, though we may well want both. Exercise 1.1 asks you to make numerical measurements of the traffic flow under various versions of the Nagel-Schreckenberg model.

1.2 Example: interpoint distances

In this section we give an example where Monte Carlo methods are used to estimate a number. The specific example we will use is one where we already know the answer. That lets us check our solution. Of course, most of this book is about problems where we don't already know the answer

Our example here is the average distance between randomly chosen points in a region. Such average distances come up in many settings. For biologists, the energy that a bird has to spend to guard its territory depends on the average distance from its nest to points in that area. In wireless networks, the probability that a distributed network is fully connected depends on average distances between communicating nodes. The average distances from dwellings to fire stations or doctors are also of interest.

To focus on essentials, suppose that $\boldsymbol{X} = (X_1, X_2)$ and $\boldsymbol{Z} = (Z_1, Z_2)$ are independent and uniformly distributed random points in a finite rectangle $R = [0, a] \times [0, b]$. Let $Y = d(\boldsymbol{X}, \boldsymbol{Z}) = \sqrt{(X_1 - Z_1)^2 + (X_2 - Z_2)^2}$ be the Euclidean distance between $\boldsymbol{X}$ and $\boldsymbol{Z}$. We can approximate the expected value of $d(\boldsymbol{X}, \boldsymbol{Z})$ by sampling pairs of points $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ for $i = 1, \dots, n$ and taking the average

$$\frac{1}{n}\sum_{i=1}^{n} d(\boldsymbol{X}_i, \boldsymbol{Z}_i). \tag{1.2}$$

Perhaps surprisingly, the exact value of $\mathbb{E}(d(\boldsymbol{X}, \boldsymbol{Z}))$ is already known in closed form, and we'll use it to see how well Monte Carlo performs. Ghosh (1951) shows that the expected distance is

$$\begin{aligned} G(a,b) = \frac{1}{15}\left[\frac{a^3}{b^2} + \frac{b^3}{a^2} + \sqrt{a^2+b^2}\left(3 - \frac{a^2}{b^2} - \frac{b^2}{a^2}\right)\right] \\ + \frac{1}{6}\left[\frac{b^2}{a}\operatorname{arccosh}\left(\frac{\sqrt{a^2+b^2}}{b}\right) + \frac{a^2}{b}\operatorname{arccosh}\left(\frac{\sqrt{a^2+b^2}}{a}\right)\right], \end{aligned} \tag{1.3}$$

where

$$\operatorname{arccosh}(t) = \log\bigl(t + \sqrt{t^2-1}\,\bigr)$$

is the upper branch of the hyperbolic arc cosine function.

For many problems, closed form expressions like equation (1.3) simply don't exist. When they do, they can be very hard to derive.

Exact solutions, like equation (1.3), are also very brittle. If we make a small change in the problem, then it may become impossible to get an exact solution. In an application, we might need to study a non-rectangular region, or use a non-uniform distribution on the region of interest, measure distance in terms of travel time, or even take points from a grid of data, such as street intersections in California. Changes like this can completely wreck a closed form solution. The situation is much more favorable for Monte Carlo. We can make a series of small changes in the way that the points are sampled or in the way that the distance is measured. With each change, the problem we're solving more closely fits the problem we set out to solve.

Consider $\boldsymbol{X}$ and $\boldsymbol{Z}$ independent and uniformly distributed in the rectangle $[0,1] \times [0,3/5]$. Equation (1.3) gives $\mathbb{E}(d(\boldsymbol{X},\boldsymbol{Y})) \doteq 0.4239$ where $\doteq$ denotes numerical rounding. It takes very little time to generate $n = 10{,}000$ pairs of points $\boldsymbol{X}_i, \boldsymbol{Z}_i \in [0,1] \times [0,3/5]$, for $1 \leqslant i \leqslant 10{,}000$. Doing this simulation one time and using equation (1.2) gave the estimate $\widehat{\mathbb{E}}(d) \doteq 0.4227$. The relative error in this estimate is $|\widehat{\mathbb{E}}(d) - \mathbb{E}(d)|/\mathbb{E}(d) \doteq 0.0027$.

In this case a small error was obtained from only 10,000 evaluations. It also takes very little time to program this simulation in a computing environment such as R or Matlab, among others. Simple Monte Carlo worked very well in this example. It was not just a lucky set of random numbers either. We defer an analysis to Chapter 2. There we will look closely at the accuracy of simple Monte Carlo, how it compares to competing methods, when it fails, and how we can estimate the error from the same sample values we use to estimate $\mathbb{E}(d)$.

1.3 Notation

Much of the notation in this book is introduced at the point where it is needed. Here we present some notational conventions that are used throughout the book, including the previous examples. The sets of real numbers and integers are denoted by $\mathbb{R}$ and $\mathbb{Z}$ respectively. Sets $\mathbb{R}^d$ and $\mathbb{Z}^d$ are then d-tuples of reals or integers respectively. This means, without always saying it, that d is an integer and $d \geqslant 1$. Whenever $d = \infty$ or even $d = 0$ is necessary, there will be a remark.

A d-tuple of real numbers is denoted in bold type, such as $\boldsymbol{x}$, $\boldsymbol{y}$, or $\boldsymbol{z}$. We write $\boldsymbol{x} = (x_1, \ldots, x_d)$ to show the components x_j of $\boldsymbol{x}$. We will call these tuples vectors because that term is more familiar. But unless $\boldsymbol{x}$ takes part in vector or matrix products we don't distinguish between the vector $\boldsymbol{x}$ and its transpose $\boldsymbol{x}^{\mathsf{T}}$. In most instances $\boldsymbol{x}$ is simply a list of d numbers used as inputs to a function. Thus, we can concatenate $\boldsymbol{x}$ and $\boldsymbol{y}$ by writing $(\boldsymbol{x}, \boldsymbol{y})$ instead of the more cumbersome $(\boldsymbol{x}^{\mathsf{T}}, \boldsymbol{y}^{\mathsf{T}})^{\mathsf{T}}$. When $d = 1$, we could equally well use $\boldsymbol{x}$ or x.

A finite sequence of vectors is denoted $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, and an infinite sequence by $\boldsymbol{x}_i$ for $i \geqslant 1$. The components of $\boldsymbol{x}_i$ are denoted x_{ij}. Thus x_7 is the seventh element of a generic vector $\boldsymbol{x}$ while $\boldsymbol{x}_7$ is the seventh vector in a sequence.

Random numbers and vectors are usually denoted by capital letters. Then, for example, $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})$ is the probability that the random vector $\boldsymbol{X} \in \mathbb{R}^d$ happens to equal the specific (nonrandom) vector $\boldsymbol{x} \in \mathbb{R}^d$. A typical Monte Carlo estimate is $(1/n)\sum_{i=1}^n f(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in \mathbb{R}^d$. It is the observed outcome of $(1/n)\sum_{i=1}^n f(\boldsymbol{X}_i)$ for random $\boldsymbol{X}_i \in \mathbb{R}^d$. We will study the mean and variance of Monte Carlo estimates, meaning the latter expression.

We write $A \in \mathbb{R}^{m\times n}$ when A is an m by n matrix of real numbers. Similarly a matrix of nonnegative numbers or binary values is an element of $[0,\infty)^{m\times n}$ or $\{0,1\}^{m\times n}$, respectively. The elements of A are A_{ij} for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$. We rarely need sequences of matrices and rarely need to compare random matrices to their observed values. Notation for these special uses is spelled out where they arise. Also, Greek letters will not necessarily be capitalized or printed in bold.

The mean and variance of a random quantity $X \in \mathbb{R}$ are written $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ respectively. The covariance of real random variables X and Y is $\mathrm{Cov}(X,Y)$. When $\boldsymbol{X} \in \mathbb{R}^d$ then $\mathbb{E}(\boldsymbol{X}) \in \mathbb{R}^d$ and $\mathrm{Var}(\boldsymbol{X}) \in \mathbb{R}^{d\times d}$. The ij element of $\mathrm{Var}(\boldsymbol{X})$ is $\mathrm{Cov}(X_i, X_j)$.

1.4 Outline of the book

The examples in this chapter used what is called simple Monte Carlo, or sometimes crude Monte Carlo. Chapter 2 explains how simple Monte Carlo works in detail. It is mainly focussed on how accurate simple Monte Carlo is. It also considers what might go wrong with simple Monte Carlo.

The next block of chapters describes what we have to know in order to use simple Monte Carlo. To begin with, we need a source of randomness. Chapter 3 describes how to simulate independent random samples from the $\mathbf{U}(0,1)$ distribution, or rather, how to make effective use of a random number generator. Chapter 4 describes how to turn $\mathbf{U}(0,1)$ random variables into non-uniform ones such as normal or Poisson random variables. That chapter covers most of the important distributions that come up in practice and, should the need arise for a brand new distribution, the general methods there can be applied. Random vectors are more complicated than random numbers because of the dependence properties linking their components. Chapter 5 shows how to sample random vectors, as well as some other random objects like rotations and permutations. The final chapter of this block is Chapter 6 which describes how to sample random processes. We need such methods, for example, when what we're sampling evolves in time and there is no fixed bound on the number of steps that have to be simulated.

Monte Carlo methods are usually presented as estimates of averages which in turn are integrals. Sometimes we can get a better answer using classical numerical integration instead of Monte Carlo. A few of the most useful quadrature methods are described in Chapter 7. In special settings we may use them instead of, or in combination with Monte Carlo.

Having worked out how to use Monte Carlo sampling, the next step is to

see how to improve upon it. Monte Carlo can be very slow to converge on some problems. Chapter 8 describes basic variance reduction techniques, including antithetic sampling, control variates, and stratification, that can give rise to faster convergence than crude Monte Carlo. An entire chapter, Chapter 9, is devoted to importance sampling. This variance reduction method is much harder to use than the others. Sometimes importance sampling brings enormous variance reductions and in other circumstances it delivers an estimate with infinite variance. Some more advanced variance reduction methods are described in Chapter 10.

There are problems that are too hard to solve by simple Monte Carlo even with all the variance reduction methods at our disposal. There are two principal sources of this difficulty. Sometimes we cannot find any practical way to make independent samples of the random inputs we need. In other settings, we can draw the samples, but the resulting Monte Carlo estimate is still not accurate enough.

Markov Chain Monte Carlo (MCMC) methods have been developed to address the first of these problems. Instead of sampling independent points, it samples from a Markov chain whose limiting distribution is the one we want. Chapter 11 presents basic theory of Markov Chain Monte Carlo, focussing on the Metropolis-Hastings algorithm. Chapter 12 describes the Gibbs sampler.

MCMC greatly expands the range of problems that Monte Carlo methods can handle. We can trace the roots of MCMC to the ideas used in generating random variables, particularly acceptance rejection sampling.

The second difficulty for Monte Carlo methods is that they can be slow to converge. We will see in Chapter 2 that the error typically decreases proportionally to $1/\sqrt{n}$ when we use n function evaluations. Quasi-Monte Carlo (QMC) and related methods improve the accuracy of Monte Carlo. The methods grow out of variance reduction methods, especially stratification. Quasi-Monte Carlo is introduced in Chapter 15 which focuses on digital net constructions. Chapter 16 presents lattice rules. QMC is deterministic and error estimation is difficult for it. Chapter 17 presents randomized quasi-Monte Carlo (RQMC) methods. These retain the accuracy of QMC while allowing independent replication for error estimation. In certain circumstances RQMC is even more accurate than QMC.

Chapter end notes

History of Monte Carlo

The Monte Carlo method has a long history. In statistics it was called 'model sampling' and used to verify the properties of estimates by mimicking the settings for which they were designed. W. S. Gosset, writing as Student (1908) derived what is now called Student's t distribution. Before finding his analytic result, he did some simulations, using height and left middle finger measurements from 3000 criminals as written on pieces of cardboard. Tippett (1927) is

an early source of numbers to use as if they were random in sampling.

Sampling was also used by physicists. Hammersley and Handscomb (1964) describe some computations done by Kelvin (1901) on the Boltzmann equation. There is more history in Kalos and Whitlock (2008) including computations made by Fermi in the 1930s. An even earlier idea was the famous Buffon needle method for estimating π by throwing needles randomly on a wooden floor and counting the fraction of needles that touch the line between two planks.

Monte Carlo sampling became far more prominent in the 1940s and early 1950s. It was used to solve problems in physics related to atomic weapons. The name itself is from this era, taken from the famous casino located in Monte Carlo. Many of the problems studied had a deterministic origin. By now it is standard to use random sampling on problems stated deterministically but early on that was a major innovation, and was even considered to be part of the definition of a Monte Carlo method.

There are numerous landmark papers in which the Monte Carlo method catches on and becomes widely used for a new class of problems. Here are some examples. Metropolis et al. (1953) present the Metropolis algorithm, the first Markov chain Monte Carlo method, for studying the relative positions of atoms. Tocher and Owen (1960) describe the GSP software (since superceded) for discrete event simulation of queues and industrial processes. Boyle (1977) shows how to use Monte Carlo methods to value financial options. Gillespie (1977) uses Monte Carlo simulation for chemical reactions in which the number of molecules is so small that differential equations are not accurate enough to describe them. Efron's (1979) bootstrap uses Monte Carlo sampling to give statistical answers with few distributional assumptions. Kirkpatrick et al. (1983) introduce simulated annealing, a Monte Carlo method for optimizing very non-smooth functions. Kajiya (1988) introduces a Monte Carlo method called path tracing for graphical rendering. Tanner and Wong (1987) use Monte Carlo algorithms to cope with problems of missing data. The realization that Markov chain Monte Carlo could be transformative for Bayesian statistical problems can be traced to Geman and Geman (1984) and Gelfand and Smith (1990) among others.

There are undoubtedly more major milestones that could be added to the list above and most of those ideas had precursors. Additional references are given in other Chapters, though not every work could be cited.

Quasi-Monte Carlo is approximately as old as Monte Carlo. The term itself was coined by Richtmyer (1952) who thought that Monte Carlo would be improved by using sequences with better uniformity than truly random sequences would have. The measurement of uniformity of sequences goes back to Weyl (1914, 1916).

Traffic modeling

The Nagel-Schreckenberg model was proposed in Nagel and Schreckenberg (1992). A comprehensive survey of traffic modeling approaches based on physics appears

in Chowdhury et al. (2000). Traffic appears as particles in some of the models and as a fluid in some others.

The description of the Nagel-Schreckenberg model in Chapter 1.1 left out the initial conditions. Here is how the data were sampled. First the N cars were placed into the M slots by sampling N values from 0 to $M-1$ without replacement. Methods for doing this are described in §5.11. Such sampling is already built-in to many computing environments. The N cars are started with initial velocity $v = 0$. Then 2500 updates were run in a 'burn-in' period. Then the 1000 time periods shown in the figure were computed.

Technical Oscars

Ken Perlin won a technical Oscar in 1997 for his work on 'solid noise' also called Perlin noise. That work was featured in the 1982 movie Tron. In 2003, Thomas Driemeyer and his team at mental images won a technical Oscar for Mental Ray rendering software. Mental Ray uses quasi-Monte Carlo.

Random distances

Ghosh (1951) finds more than just the mean distance. He finds the probability density function of the distance d, as well as $\mathbb{E}(d^k)$ for $k = 1, 2, 3, 4$. For the density function, three different formulas are required, one for each of the ranges $[0, b)$, $[b, a)$, $[a, \sqrt{a^2 + b^2}]$.

Much of Ghosh's paper is concerned with the two rectangle problem. The points $\boldsymbol{X}$ and $\boldsymbol{Z}$ are sampled from two different and possibly overlapping rectangles, with parallel sides. The density function of $d(\boldsymbol{X}, \boldsymbol{Z})$ takes 15 different formulas on 15 different ranges. Instead of handling all of these cases, Ghosh considers special versions of the problem such as the one in which the rectangles are congruent squares touching at a corner or at a side. Marsaglia et al. (1990) give a survey of random distances in rectangles, including many from before Ghosh. They include figures of the probability density of those distances for rectangles with varying relative sizes and amounts of overlap.

The average distance between two independent and uniformly distributed points in $[0, 1]^3$ is known as the Robbins constant after Robbins (1978). It is

$$\begin{aligned}\Delta(3) &= \frac{1}{105}\Big[4 + 17\sqrt{2} - 6\sqrt{3} + 21\log(1+\sqrt{2}) + 42\log(2+\sqrt{3}) - 7\pi\Big] \\ &\doteq 0.66160. \end{aligned} \tag{1.4}$$

While the mean is known, the probability density of the three dimensional random distance was not known as of 2005 (Weisstein, 2005).

Exercises

These exercises require the use of random number generators. The basics of random number generators are described in Chapter 3. The exercises in this

chapter do not require sophisticated uses of random numbers and so can be approached before reading that chapter.

1.1. For this exercise, implement the Nagel-Schreckenberg traffic model described in §1.1. There are M slots on a circular road and $k < M$ of them are occupied by cars each with velocity 0. To start the simulation, you may choose those k occupied slots to be approximately equispaced. Then run the simulation for a burn-in period of B time steps so as to largely forget the initial configuration. Start with the k cars nearly equispaced.

a) For $M = B = 1000$ and $v_{\max} = 5$ and $k = 50$ and $p = 1/3$ run the simulation for 1000 updates after burn-in. Make a flow trace image analogous to Figure 1.1 for this example. Report the total distance traveled by the k cars in their last 1000 steps (not counting their burn-in distance).

b) Repeat the previous item for $k = 55$ through $k = 500$ by steps of 5. Plot the total distance traveled by all the cars versus k. This figure, plotting flow versus density, is known as the fundamental diagram. Roughly what value of k gives the best flow? Select two of these new values of k and plot their flow traces. Explain why you chose these values of k and what they show about the traffic model.

c) Repeat the simulations in the part b) four more times. Add the four new results to the previous plot. Make a second plot with the mean flow at each value of k along with the upper and lower approximate 99% confidence limits for the mean flow at that value of k. Roughly what range of k values has the maximum flow rate?

d) Create 5 more fundamental diagram plots, but this time start the cars in positions 0 through $k-1$. Does it make any difference after the 1000 step burn-in?

We have not yet studied how to synchronize multiple simulations. For this problem take every random number used to be independent of every other one.

1.2. High dimensional geometry tells us that most of the volume in the hypercube $[0,1]^d$ is near the boundary, when d is large. Specifically, letting

$$d_B(\boldsymbol{x}) = \min_{1\leqslant j\leqslant d} \min(x_j, 1-x_j)$$

be the distance from $\boldsymbol{x}$ to the boundary, and $A_{\epsilon,d} = \{\boldsymbol{x} \in [0,1]^d \mid d_B(\boldsymbol{x}) \leqslant \epsilon\}$ for $0 < \epsilon < 1/2$ and $d \geqslant 1$, then $\mathbf{vol}(A_{\epsilon,d}) = 1 - (1-2\epsilon)^d$.

On the other hand, for large d, the central limit theorem tells us that most of the points $\boldsymbol{x} \in [0,1]^d$ are near a hyperplane $\{\boldsymbol{x} \mid \sum_{j=1}^d x_j = d/2\}$, which passes through the center of the cube. Letting

$$d_C(\boldsymbol{x}) = \frac{1}{\sqrt{d}}\left|\sum_{j=1}^d (x_j - 1/2)\right|$$

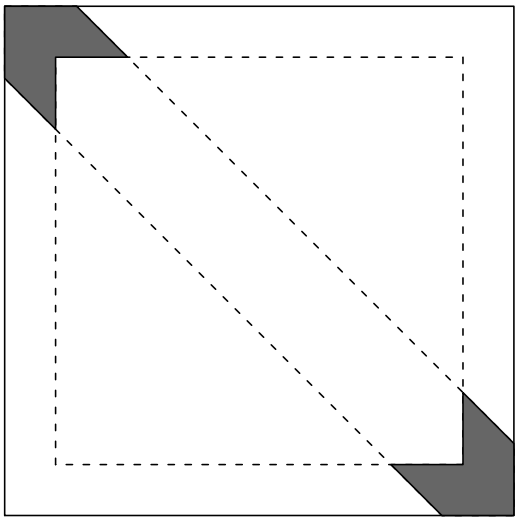

Figure 1.2: This figure illustrates Exercise 1.2 for the case $d = 2$ and $\epsilon = 0.1$. The boundary is the unit square. Points outside the dashed square are within ϵ of the boundary. Points between the two parallel dashed lines are within ϵ of the center plane $\sum_j x_j = d/2$. The shaded region is within ϵ of both the center plane and the boundary. In high dimensions, most of the volume would be shaded.

be the distance of $\boldsymbol{x}$ to the center plane, $B_{\epsilon,d} = \{\boldsymbol{x} \in [0,1]^d \mid |d_C(\boldsymbol{x})| \leqslant \epsilon\}$ has volume nearly $1 - 2\Phi(-\epsilon\sqrt{12d})$, for large d.

The shaded region in Figure 1.2 is $A_{\epsilon,2} \cap B_{\epsilon,2}$, and so for small d the intersection of these sets is not large. Use Monte Carlo to estimate $\mathbf{vol}(A_{\epsilon,d} \cap B_{\epsilon,d})$ for $\epsilon = 0.1$ and $d \in \{2, 3, 4, 5, 10, 20\}$.

1.3. For this problem we reconsider the average travel time between points in the rectangle $R \equiv [0,1] \times [0,0.6]$. But now, travel can be greatly accelerated by an expressway over the circle with center 0 and radius 1/2. To get from point $\boldsymbol{a} = (a_1, a_2)$ to point $\boldsymbol{b} = (b_1, b_2)$, both in R, we could use the direct route from $\boldsymbol{a}$ to $\boldsymbol{b}$, which takes time

$$t_D = ((a_1 - b_1)^2 + (a_2 - b_2)^2)^{1/2}.$$

Or we could travel from $\boldsymbol{a}$ to a point $\widetilde{\boldsymbol{a}} = (\widetilde{a}_1, \widetilde{a}_2)$ in the express circle $\mathcal{C} = \{(c_1, c_2) \in R \mid c_1^2 + c_2^2 = 1/2\}$ then from $\widetilde{\boldsymbol{a}}$ to $\widetilde{\boldsymbol{b}} = (b_1', b_2') \in \mathcal{C}$, and then from $\widetilde{\boldsymbol{b}}$ to $\boldsymbol{b}$. That route takes a total time of

$$t_E = \big((a_1 - \widetilde{a}_1)^2 + (a_2 - \widetilde{a}_2)^2\big)^{1/2} + \big((b_1 - \widetilde{b}_1)^2 + (b_2 - \widetilde{b}_2)^2\big)^{1/2},$$

because travel time from $\widetilde{\boldsymbol{a}}$ to $\widetilde{\boldsymbol{b}}$ on the express circle is negligible. Naturally we

pick $\widetilde{\boldsymbol{a}}$ as close to $\boldsymbol{a}$ as possible and $\widetilde{\boldsymbol{b}}$ as close to $\boldsymbol{b}$ as possible to minimize total time. The actual travel time is $t = \min(t_D, t_E)$.

a) Find and report the travel times from point $\boldsymbol{a}$ to point $\boldsymbol{b}$ for the points given in Table 1.1. **Hint:** It is a good idea to make a plot of R, $\mathcal{C}$ and the points $\boldsymbol{a}$, $\boldsymbol{b}$, to provide a visual check on your answers.

$\boldsymbol{a}$		$\boldsymbol{b}$	
a_1	a_2	b_1	b_2
0.45	0.05	0.05	0.45
0.04	0.13	0.18	0.32
0.56	0.49	0.16	0.50
0.20	0.30	0.95	0.40
0.54	0.30	0.40	0.59
0.00	0.00	0.60	0.45

Table 1.1: This table has points in $[0, 1] \times [0, 0.6]$ used in Exercise 1.3.

b) Suppose that you know that neither $\boldsymbol{a} = (a_1, a_2)$ nor $\boldsymbol{b} = (b_1, b_2)$ is $(0, 0)$. Give a simple expression for the travel time from $\boldsymbol{a}$ to $\boldsymbol{b}$ in terms of a_1, a_2, b_1, and b_2.

c) What is the average value of t when $\boldsymbol{a}$ and $\boldsymbol{b}$ are independent random locations uniformly distributed within the rectangle R?

2

Simple Monte Carlo

The examples in Chapter 1 used ***simple Monte Carlo***. That is a direct simulation of the problem of interest. Simple Monte Carlo is often called ***crude Monte Carlo*** to distinguish it from more sophisticated methods. Despite these mildly pejorative names, simple Monte Carlo is often the appropriate method.

In this chapter we look at further issues that come up in simple Monte Carlo. In simple Monte Carlo our goal is to estimate a population expectation by the corresponding sample expectation. That problem has been very thoroughly studied in probability and statistics. This chapter shows how to apply results from probability and statistics to simple Monte Carlo. Using the laws of large numbers and the central limit theorem, we can see how accurate simple Monte Carlo is. We also derive confidence intervals for a sample mean, using the sample data values themselves.

The basics of simple Monte Carlo are given in §2.1 to §2.3. To complete the chapter, we look at specialized topics in simple Monte Carlo. Some readers might prefer to skip those topics until they face the problems described. Those topics include using simple Monte Carlo to estimate probabilities, quantiles, ratios, and other smooth functions of means. We give confidence interval methods for each problem. We also look at very heavy-tailed settings that can cause simple Monte Carlo to work poorly, or in extreme cases, fail completely.

2.1 Accuracy of simple Monte Carlo

In a simple Monte Carlo problem we express the quantity we want to know as the expected value of a random variable Y, such as $\mu = \mathbb{E}(Y)$. Then we generate values $Y_1, \dots, Y_n$ independently and randomly from the distribution of Y and

take their average

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{2.1}$$

as our estimate of μ.

In practice, there is usually a bit more to the story. Commonly $Y = f(\boldsymbol{X})$ where the random variable $\boldsymbol{X} \in \mathcal{D} \subseteq \mathbb{R}^d$ has a probability density function $p(\boldsymbol{x})$, and f is a real-valued function defined over $\mathcal{D}$. Then $\mu = \int_{\mathcal{D}} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. For some problems it is easier to work with expectations while for other tasks it is simpler to work directly with the integrals. In still other settings $\boldsymbol{X}$ is a discrete random variable with a probability mass function that we also call p. The input $\boldsymbol{X}$ need not even be a point in Euclidean space at all. It could be the path taken by a wandering particle or it could be an image. But so long as $Y = f(\boldsymbol{X})$ is a quantity that can be averaged, such as a real number or vector, we can apply simple Monte Carlo.

The primary justification for simple Monte Carlo is through the laws of large numbers. Let Y be a random variable for which $\mu = \mathbb{E}(Y)$ exists, and suppose that $Y_1, \dots, Y_n$ are independent and identically distributed with the same distribution as Y. Then under the ***weak law of large numbers***,

$$\lim_{n\to\infty} \mathbb{P}\big(|\hat{\mu}_n - \mu| \leqslant \epsilon\big) = 1, \tag{2.2}$$

holds for any $\epsilon > 0$. The weak law tells us that our chance of missing by more than ϵ goes to zero. The ***strong law of large numbers*** tells us a bit more. The absolute error $|\hat{\mu}_n - \mu|$ will eventually get below ϵ and then stay there forever:

$$\mathbb{P}\Big(\lim_{n\to\infty} |\hat{\mu}_n - \mu| = 0\Big) = 1. \tag{2.3}$$

We had to assume that μ exists. This is an extremely mild assumption. If μ did not exist why would we be trying to estimate it? In §2.8 we explore what happens when this and other usually mild assumptions fail, and how we might detect such problems.

While both laws of large numbers tell us that Monte Carlo will eventually produce an error as small as we like, neither tells us how large n has to be for this to happen. They also don't say for a given sample $Y_1, \dots, Y_n$ whether the error is likely to be small.

The situation improves markedly when Y has a finite variance. Suppose that $\mathrm{Var}(Y) = \sigma^2 < \infty$. In IID sampling, $\hat{\mu}_n$ is a random variable and it has its own mean and variance. The mean of $\hat{\mu}_n$ is

$$\mathbb{E}(\hat{\mu}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i) = \mu. \tag{2.4}$$

Because the expected value of $\hat{\mu}_n$ is equal to μ we say that simple Monte Carlo is ***unbiased***. The variance of $\hat{\mu}_n$ is

$$\mathbb{E}((\hat{\mu}_n - \mu)^2) = \frac{\sigma^2}{n}, \tag{2.5}$$

by elementary manipulations.

While it is intuitively obvious that the answer should get worse with increased variance and better with increased sample size, equation (2.5) gives us the exact rate of exchange. The root mean squared error (RMSE) of $\hat{\mu}_n$ is $\sqrt{\mathbb{E}((\hat{\mu}_n - \mu)^2)} = \sigma/\sqrt{n}$. To emphasize that the error is of order $n^{-1/2}$ and de-emphasize σ, we write RMSE $= O(n^{-1/2})$ as $n \to \infty$. The $O(\cdot)$ notation is described on page 35 of the chapter end notes. To get one more decimal digit of accuracy is like asking for an RMSE one tenth as large, and that requires a 100-fold increase in computation. To get three more digits of accuracy requires one million times as much computation. It is clear that simple Monte Carlo computation is poorly suited for problems that must be answered with high precision.

Equation (2.5) also shows that if we can change the problem in some way that reduces σ^2 by a factor of two while leaving μ unchanged, then we gain just as much as we would by doubling n. If we can recode the function to make it twice as fast, or switch to a computer that is twice as fast, then we make the same gain as we would get by cutting σ^2 in two. The economics of the $\sigma/\sqrt{n}$ error rate also work in reverse. If raising n from n_1 to n_2 only makes our accuracy a little better, then reducing n from n_2 to n_1 must only make our accuracy a little worse. We might well decide that software that runs slower, reducing n for a given time budget, is worthwhile if it provides some other benefit, such as a more rapid programming.

The $n^{-1/2}$ rate is very slow compared to the rate seen in many numerical problems. For example, when $d = 1$, then Simpson's method can integrate a function with an error that is $O(n^{-4})$, when the integrand has a continuous fourth derivative, as described in §7.2. Simpson's rule with n points is then about as accurate as Monte Carlo with An^8 points, where A depends on the relationship between the fourth derivative of the integrand and its variance.

While low accuracy cannot be considered a strength, it is not always a severe weakness. Sometimes we only need a rough estimate of μ in order to decide what action to take. At other times, there are model errors that are probably larger than the Monte Carlo errors. Our models usually make some idealized assumptions, such as specific distributional forms. Even with the right distributional form we may employ incorrect values for parameters such as future prices or demand levels. The advantage of a Monte Carlo approach is that we can put more real world complexity into our computations than we would be able to get with closed form estimates.

A striking feature about the formula $\sigma/\sqrt{n}$ is that the dimension d of the argument $\boldsymbol{x}$ does not appear in it anywhere. In applications d can be 2 or d can be 1000 and the RMSE is still $\sigma/\sqrt{n}$. By contrast, a straightforward extension of Simpson's rule to d dimensions has an error rate of $O(n^{-4/d})$, as described in §7.4, making it useless for large d.

Another feature that does not appear in (2.5) is the smoothness of f. Competing methods like Simpson's rule that beat Monte Carlo in low dimensional smooth problems can do badly when f is not smooth. The rate in Simpson's rule requires a bounded fourth derivative for f. Simple Monte Carlo is most

competitive in high dimensional problems that are not smooth, and for which closed forms are unavailable.

2.2 Error estimation

One of the great strengths of the Monte Carlo method is that the sample values themselves can be used to get a rough idea of the error $\hat{\mu}_n - \mu$. A rough idea is usually good enough. We are usually more interested in a good estimate of μ itself than of the error.

The average squared error in Monte Carlo sampling is σ^2/n. We seldom know σ^2 but it is easy to estimate it from the sample values. The most commonly used estimates of σ^2 are

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}_n)^2, \quad \text{and,} \tag{2.6}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_n)^2. \tag{2.7}$$

Monte Carlo sampling typically usually uses such large values of n that (2.6) and (2.7) will be much closer to each other than either of them is to actual variance σ^2. The familiar motivation for (2.6) is that it is unbiased:

$$\mathbb{E}(s^2) = \sigma^2, \tag{2.8}$$

for $n \geqslant 2$. Both formulas will appear in the variance estimates that we use.

A variance estimate s^2 tells us that our error is on the order of $s/\sqrt{n}$. We know that $\hat{\mu}_n$ has mean μ and we can estimate its variance by s^2/n. From the central limit theorem (CLT), we also know that $\hat{\mu}_n - \mu$ has approximately a normal distribution with mean 0 and variance σ^2/n.

Before stating the CLT we need some notation. The normal distribution with mean 0 and variance 1, called the standard normal distribution, has probability density function

$$\varphi(z) = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}, \qquad -\infty < z < -\infty \tag{2.9}$$

and cumulative distribution function

$$\Phi(a) = \int_{-\infty}^{a} \varphi(z)\,dz, \qquad -\infty < z < -\infty. \tag{2.10}$$

The normal quantile function Φ^{-1} maps $(0,1)$ onto $\mathbb{R}$. When Z has the standard normal distribution, we write $Z \sim \mathcal{N}(0,1)$.

Theorem 2.1 (IID central limit theorem)**.** *Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random variables with mean μ and finite variance $\sigma^2 > 0$. Let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Then for all $z \in \mathbb{R}$*

$$\mathbb{P}\Big(\sqrt{n}\,\frac{\hat{\mu}_n - \mu}{\sigma} \leqslant z\Big) \to \Phi(z), \tag{2.11}$$

as $n \to \infty$.

Proof. Chung (1974). □

We will sometimes use a vector version of Theorem 2.1. If $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^d$ are IID with mean $\mu \in \mathbb{R}^d$ and finite variance-covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, then $\sqrt{n}(\bar{\boldsymbol{Y}} - \mu) \to \mathcal{N}(0, \Sigma)$ as $n \to \infty$, where $\bar{\boldsymbol{Y}} = (1/n)\sum_{i=1}^n \boldsymbol{Y}_i$ and $\mathcal{N}(0, \Sigma)$ denotes a d dimensional multivariate normal distribution with mean 0 and variance Σ. We postpone a description of the multivariate normal distribution until §5.2 where we see how to draw samples from $\mathcal{N}(\mu, \Sigma)$.

Theorem 2.1 can be used to get approximate confidence intervals for μ, but it requires that we know σ. As $n \to \infty$, $\mathbb{P}(|s - \sigma| > \epsilon) \to 0$ for any $\epsilon > 0$, by a short argument applying the weak law of large numbers to both $(1/n)\sum_{i=1}^n Y_i$ and $(1/n)\sum_{i=1}^n Y_i^2$. As a result, we can substitute s for σ. That is, under the conditions of Theorem 2.1

$$\mathbb{P}\Big(\sqrt{n}\,\frac{\hat{\mu}_n - \mu}{s} \leqslant z\Big) \to \Phi(z) \tag{2.12}$$

as $n \to \infty$. This plug-in operation can be justified by Slutsky's Theorem, which is described on page 38 of the chapter end notes.

Restating equation (2.12) we find for $\Delta > 0$ that

$$\begin{aligned}\mathbb{P}\Big(|\hat{\mu}_n - \mu| \geqslant \frac{\Delta s}{\sqrt{n}}\Big) &= \mathbb{P}\Big(\sqrt{n}\,\frac{\hat{\mu}_n - \mu}{s} \leqslant -\Delta\Big) + \mathbb{P}\Big(\sqrt{n}\,\frac{\hat{\mu}_n - \mu}{s} \geqslant \Delta\Big)\\ &\to \Phi(-\Delta) + (1 - \Phi(\Delta))\\ &= 2\Phi(-\Delta)\end{aligned}$$

by symmetry of the $\mathcal{N}(0,1)$ distribution. For a 95% confidence interval we allow a 5% chance of non-coverage, and therefore set $2\Phi(-\Delta) = 0.05$. Then $\Delta = -\Phi^{-1}(.025) = \Phi^{-1}(0.975) \doteq 1.96$, yielding the familiar 95% confidence interval

$$\hat{\mu}_n - 1.96\,\frac{s}{\sqrt{n}} \leqslant \mu \leqslant \hat{\mu}_n + 1.96\,\frac{s}{\sqrt{n}}.$$

We will write such intervals as $\hat{\mu}_n \pm 1.96 s/\sqrt{n}$. We might well prefer a 99% confidence interval, and then we have only to replace $\Phi^{-1}(0.975)$ by $\Phi^{-1}(0.995) \doteq 2.58$. To summarize, equation (2.12) justifies CLT-based approximate confidence intervals of the form

$$\begin{aligned}&\hat{\mu}_n \pm 1.96\, s/\sqrt{n}\\ &\hat{\mu}_n \pm 2.58\, s/\sqrt{n}, \quad \text{and,}\end{aligned} \tag{2.13}$$

$$\hat{\mu}_n \pm \Phi^{-1}(1-\alpha/2)s/\sqrt{n}, \tag{2.14}$$

for 95, 99, and $100(1-\alpha)$ percent confidence respectively.

In special circumstances we may want a probabilistic bound for μ. We may then use ***one-sided confidence intervals*** $(-\infty, \hat{\mu} + \Phi^{-1}(1-\alpha)s/\sqrt{n}]$ or $[\hat{\mu} + \Phi^{-1}(\alpha)s/\sqrt{n}, \infty)$, for upper or lower bounds, respectively.

Example 2.1 (Ghosh's distances)**.** The example in §1.2 had $n = 10{,}000$ sample distances between pairs of points randomly selected in $[0,1] \times [0, 0.6]$. The mean value of d was 0.4227 with $s^2 = 0.04763$. Therefore the 99% confidence interval is $0.4227 \pm 2.58\sqrt{0.04763}/\sqrt{10000}$ which gives the interval $[0.4171, 0.4284]$. As it happens, the true value $\mu = 0.4239$ is in this interval. The true variance of d is easy to compute because $\mathbb{E}(d^2)$ and $\mathbb{E}(d)$ are both known. To three significant figures, the variance σ^2 of d is 0.0469. The reasonably accurate estimate from §1.2 was not unusually lucky. It missed by about 1/2 of $\sigma/\sqrt{n}$. Also s^2 was quite close to the true σ^2.

Most of the 99% confidence intervals we will use are generalizations of (2.13) taking the form $\hat{\mu}_n \pm 2.58\sqrt{\widehat{\mathrm{Var}}(\hat{\mu}_n)}$. Here $\hat{\mu}_n$ is an unbiased or nearly unbiased estimate of μ for which a central limit theorem holds, and $\widehat{\mathrm{Var}}(\hat{\mu}_n)$ is an unbiased, or nearly unbiased estimate of $\mathrm{Var}(\hat{\mu}_n)$.

It is also very common to replace the normal quantile $\Phi^{-1}(1-\alpha/2)$ by one from Student's t distribution on $n-1$ degrees of freedom. If the sampled Y_i are normally distributed, then intervals of the form

$$\hat{\mu}_n \pm t_{(n-1)}^{1-\alpha/2}\, s/\sqrt{n} \tag{2.15}$$

have exactly $1-\alpha$ probability of containing μ. Here $t_{(n-1)}^{\eta}$ denotes the η quantile of the $t_{(n-1)}$ distribution. Monte Carlo applications usually use n so large that there is no practical difference between the t–based confidence intervals from (2.15) and confidence intervals based on (2.14). From numerical t tables, the 99% intervals with the $t_{(n)}$ distribution are about $1 + 1.9/n$ times as wide as those based on $\mathcal{N}(0,1)$ for $n \geqslant 1000$.

For nearly normal Y_i and small n, such as $n \leqslant 20$, using equation (2.15) makes a difference. There really are Monte Carlo settings where $n \leqslant 20$ is a reasonable choice. For example each Y_i may be the result of a very complicated Monte Carlo simulation with millions of function values, for which we have no good error estimate. Repeating that whole process a small number n of times and then using (2.15) supplies an error estimate.

In Monte Carlo applications, we almost always use approximate confidence intervals, like those based on the central limit theorem. It is usual to omit the term 'approximate'. There are a few exceptions, such as estimating probabilities in §2.4, for which exact confidence intervals can be obtained.

The accuracy level of confidence intervals has also been studied. The derivation uses methods outside the scope of this book, but the results are easily stated. Typically

$$\mathbb{P}\big(|\mu - \hat{\mu}| \leqslant \Phi^{-1}(1-\alpha/2)s/\sqrt{n}\,\big) = 1 - \alpha + O(n^{-1}) \tag{2.16}$$

as $n \to \infty$. That is, when we're aiming for 99% confidence we actually get $99\% + O(1/n)$. Equation (2.16) requires $\mathbb{E}(Y^4) < \infty$ and it does not apply if Y is constrained to a one dimensional shifted lattice $\{a + bx \mid x \in \mathbb{Z}\}$ for $a, b \in \mathbb{R}$. This high accuracy does not apply to one-sided confidence intervals. Their error level is typically $O(1/\sqrt{n})$. See Hall (1992) for these results and many more.

Equation (2.16) is remarkable. It implies that the accuracy of the confidence interval is actually better than the $O(1/\sqrt{n})$ accuracy of the Monte Carlo estimate $\hat{\mu}$ itself.

2.3 Safely computing the standard error

In order to compute CLT-based confidence intervals we need first to compute s. The formula for s is simple enough but, perhaps surprisingly, it can lead to numerical difficulties. Numerical problems may arise when n is large and when $\sigma \ll |\mu|$. Since modern computers make very large n reasonable (e.g., $n \geqslant 10^9$), the issue is worth investigating.

Formulas (2.6) and (2.7), taken literally, suggest making two passes over the Y_i values, first to compute $\hat{\mu}_n$ and then to compute s^2 or $\hat{\sigma}^2$. Making two passes requires us to either store or recompute the values Y_i. There is a natural (but numerically unstable) way to compute these error estimates in one pass. For $\hat{\sigma}^2$ we may write

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \Big(\frac{1}{n}\sum_{i=1}^{n} y_i\Big)^2 \qquad \text{(Don't compute this way!)} \tag{2.17}$$

where y_i is the sampled value of the random variable Y_i. Equation (2.17) serves for s^2 too because $s^2 = n\hat{\sigma}^2/(n-1)$. The problem with (2.17) is that it can lead to very bad roundoff errors from 'catastrophic cancellation'. When s is much smaller than $\hat{\mu}_n$, then the quantities in the subtraction may be of nearly equal magnitude and the difference may be dominated by rounding error. The result may be a computed value of s^2 that fails to have even one correct digit of accuracy. It is even possible to compute a negative value for $\hat{\sigma}^2$. At least the negative value provides a clear signal that something has gone wrong. A positive but erroneous value may be worse because it is less likely to be detected.

There is a way to obtain good numerical stability in a one-pass algorithm. Let $S_n = \sum_{i=1}^{n}(y_i - \hat{\mu}_n)^2$. Starting with $\hat{\mu}_1 = y_1$ and $S_1 = 0$, make the updates

$$\begin{aligned} \delta_i &= y_i - \hat{\mu}_{i-1} \\ \hat{\mu}_i &= \hat{\mu}_{i-1} + \frac{1}{i}\,\delta_i \\ S_i &= S_{i-1} + \frac{i-1}{i}\,\delta_i^2 \end{aligned} \tag{2.18}$$

for $i = 2, \ldots, n$. Then use $s^2 = S_n/(n-1)$ in approximate confidence intervals. See Exercise 2.3.

For very large even numbers n, the following very simple estimate is useful:

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n/2}(y_{2i} - y_{2i-1})^2. \tag{2.19}$$

It is also easy to use (2.19) in parallel computation which we may need for very large n. When we are really interested in estimating σ^2 per se, then s^2 is much better than $\widetilde{\sigma}^2$. See Exercise 2.5. But when we are primarily interested in getting a confidence interval for μ, then $\widetilde{\sigma}$, used in

$$\hat{\mu}_n \pm 2.58\widetilde{\sigma}/\sqrt{n},$$

is almost as good as s statistically, and could be much better numerically (Exercise 2.6). Exercise 2.7 considers a more radical shortcut.

2.4 Estimating probabilities

An important special case arises when the function to be averaged only has two possible values, conventionally taken to be 0 and 1, or failure and success, respectively. Let $f(\boldsymbol{x}) = 1$ for $\boldsymbol{x} \in A$ and 0 for $\boldsymbol{x} \notin A$. We write such indicator functions as $f(\boldsymbol{x}) = \mathbb{1}_{\boldsymbol{x}\in A}$ or alternatively as $f(\boldsymbol{x}) = \mathbb{1}_A(\boldsymbol{x})$. Suppose that $\boldsymbol{X}$ has probability density function p. Then $\mathbb{E}(f(\boldsymbol{X}))$ is simply $\mathbb{P}(\boldsymbol{X} \in A)$ for $\boldsymbol{X} \sim p$.

As an example, consider a wireless multi-hop network. To keep things simple, suppose that there are $m \geqslant 3$ nodes randomly distributed in the unit square. Each node can communicate directly with any other node that is within a distance r of it. A two-hop connection arises between nodes 1 and 2 when nodes 1 and 2 are farther than r apart but are both within distance r of node j for one or more $j \in \{3, \ldots, m\}$.

Figure 2.1 illustrates the issue when $m = 20$. Circles of radius $r = 0.2$ are drawn around nodes 1 and 2. In the left panel, nodes 1 and 2 are within r of each other, so there is a one-hop connection. In the middle panel, nodes 1 and 2 have a two-hop connection, but no one-hop connection. In the right panel, nodes 1 and 2 are farther than $2r$ apart, so no two-hop connection is possible.

One thing, among many, that we would like to know is the probability of a two-hop connection. This probability is an integral over $2m = 40$ dimensional space, of a binary function that is 1 when and only when there is a two-hop, but no one-hop, connection between nodes 1 and 2. The examples in Figure 2.1 were selected from 10,000 independent replications of this problem. The best connection between nodes 1 and 2 turned out to be a two-hop connection in 650 of the replications. That is, the estimated probability is $650/10000 = 0.065$.

For a binary random variable $Y \in \{0, 1\}$, the variance is completely determined by the mean. Suppose that $Y = 1$ with probability $p \in [0, 1]$. Then $\mathbb{E}(Y) = p$ and $\mathbb{E}(Y^2) = \mathbb{E}(Y) = p$ too, because $Y^2 = Y$ for a binary variable. Therefore $\sigma^2 = p - p^2 = p(1 - p)$. In the binary case we prefer to write the

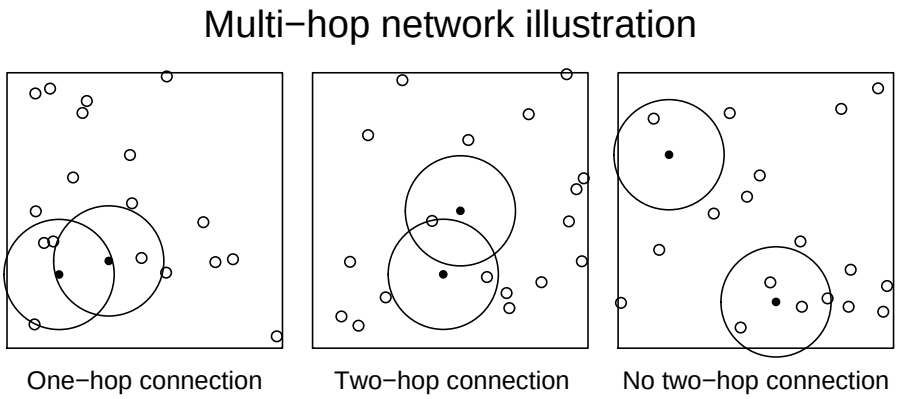

Figure 2.1: Each panel has 20 points in the unit square. There is a one-hop connection between the circled solid points in the left panel. There is no one-hop connection between solid points in the middle panel, but there is (just barely) a two-hop connection. There is no two-hop connection between the solid points in the right panel.

sample value $\hat{\mu}_n$ as $\hat{p}_n$. The CLT-based 99% confidence interval for p is usually taken to be

$$\hat{p}_n \pm 2.58\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}. \tag{2.20}$$

For the example of this section we get

$$0.065 \pm 2.58\sqrt{\frac{0.065 \times 0.945}{10000}} \doteq 0.065 \pm 0.0064.$$

If we plug the binary valued Y_i into (2.13) the result is not quite (2.20). To force a match we could use $n-1$ instead of n in the denominator of (2.20), but it is more common to just accept a small difference.

The binary case has one special challenge. Sometimes we get no 1's in the data. In that case equation (2.20) gives us a confidence interval from 0 to 0. Even if we thought that $\hat{p}_n = 0$ was a reasonable best estimate of p, we should hesitate to take 0 as an upper confidence limit. The chance of getting no successes in n tries is $(1-p)^n$. Supposing that this chance was not below one percent, that is $(1-p)^n \geqslant 0.01$, then we find $p \leqslant 1 - 0.01^{1/n}$. For large n, using a Taylor approximation,

$$1 - 0.01^{1/n} = 1 - e^{\frac{1}{n}\log(0.01)} \doteq 1 - \Big(1 + \frac{1}{n}\log(0.01)\Big) = \frac{\log(100)}{n} \doteq \frac{4.605}{n}.$$

When $\hat{p}_n = 0$, a reasonable 99% confidence interval for p is $[0, 4.605/n]$. A similar argument using $(1-p)^n \leqslant 0.05$ gives the well known approximate 95% confidence interval $[0, 3/n]$, when 0 successes are seen in n trials.

In the binary case, we can get exact confidence intervals. Let $T = n\hat{p}$ be the total number of successes in n trials. Then a counting argument shows that T has the binomial distribution

$$\mathbb{P}(T = t) = \binom{n}{t} p^t (1-p)^{n-t},$$

with parameters n and p. An exact 99% confidence interval $p_L \leqslant p \leqslant p_U$ can be found, when $T = t$, by solving

$$0.005 = \sum_{i=t}^{n} \binom{n}{i} p_L^i (1-p_L)^{n-i}, \qquad \text{and,} \tag{2.21}$$

$$0.005 = \sum_{i=0}^{t} \binom{n}{i} p_U^i (1-p_U)^{n-i}, \tag{2.22}$$

except that $p_L = 0$ when $T = 0$ and $p_U = 0$ when $T = n$. Many computing environments include functions to compute the right hand sides of (2.21) and (2.22). We may still need a bisection or other search method to find p_L and p_U.

A pragmatic approach to handling binomial confidence intervals is to add a pseudocount of a_0 failures and a_1 successes and then treat the data as if $T + a_1$ successes and $(n - T) + a_0$ failures were obtained in $n + a_0 + a_1$ trials. Agresti (2002) recommends taking $a_0 = a_1 = \Phi^{-1}(1-\alpha/2)^2/2$ for $100(1-\alpha)$ confidence. For 99% confidence $\alpha = 0.01$ and so $a_0 = a_1 \doteq 3.317$. Then the confidence interval is

$$\widetilde{p}_n \pm 2.58\sqrt{\frac{\widetilde{p}_n(1-\widetilde{p}_n)}{n}} \tag{2.23}$$

where

$$\widetilde{p}_n = \frac{n\hat{p}_n + 3.317}{n + 6.634}.$$

Table 2.1 shows the results from the three 99% confidence regions applied to the multi-hop wireless example. There is little difference between the methods on this problem. Big differences can arise, but they are less likely to do so when n is as large as 10,000 and p is not close to 0 or 1.

When the number $n\hat{p}_n$ of successes is close to 0 or to n then the intervals can differ. Both the Agresti and the CLT intervals can have negative lower limits. In practice we replace those by 0. Similarly, upper limits above 1 can be replaced by 1.

The really hard cases arise for rare events. If we chose to study a very small radius in the two-hop problem, then we could well find only 0, 1 or a handful of successes, even in millions of trials. From one point of view we might be satisfied to get 0 successes in n trials and then have 99% confidence that $p \leqslant 4.605/n$. For $n = 10^6$ we would get a very narrow interval for p. But it is more likely

	Lower	Upper
CLT	0.05865	0.07135
Exact	0.05881	0.07161
Agresti	0.05893	0.07165

Table 2.1: For the two-hop problem there were 650 successes in 10,000 trials. This table shows three different 99% confidence intervals discussed in the text as applied to this example.

that we would want to know a small p to within some reasonably small relative error. The width of the Agresti interval, for example, is

$$2 \times 2.58 \times \sqrt{\widetilde{p}_n(1-\widetilde{p}_n)/n} \doteq 5.16 \times \sqrt{\widetilde{p}_n/n}$$

when $\widetilde{p}_n$ is near 0. The relative error is then about $5.16\sqrt{\widetilde{p}_n/n}/\widetilde{p}_n = 5.16/\sqrt{n\widetilde{p}_n}$, which will usually be large when p is small. To get a relative error below Δ requires $n \geqslant 26.6/(\Delta^2\widetilde{p}_n)$. The required sample size grows inversely with the success probability becoming prohibitive for rare events. Methods to handle rare events are described in Chapter 9 on importance sampling.

2.5 Estimating quantiles

While most simple Monte Carlo projects are aimed at estimating averages, there are lots of important problems where we want to estimate a median, or perhaps more commonly, an extreme quantile of a distribution.

In financial ***value at risk*** problems, the random variable $Y \sim F$ may represent the future value of our portolio. The 0.01 quantile, $Q^{0.01} = Q^{0.01}(F)$ is defined by

$$\mathbb{P}(Y \leqslant Q^{0.01}) = 0.01.$$

From $Q^{0.01}$ we can judge how much we lose in a not quite worst-case scenario. Similarly, the reliability of power plants, wind resistances of buildings, and exposures to environmental toxins are appropriately measured via quantiles. Monte Carlo methods for these problems are referred to as ***probabilistic risk assessment***. Likewise, in many industries ***quality of service*** may be based on a quantile of the customers' waiting times, among other measures.

In statistical inference, a task is to find the cutoff value at which a test statistic attains 0.05 significance. Let $T \geqslant 0$ be a random variable for which larger values describe greater departures from a simplified model, called the null hypothesis. The 0.95 quantile $Q^{0.95}$ defined by $\mathbb{P}(T \leqslant Q^{0.95}) = 0.95$ is a critical dividing line. If we see $T > Q^{0.95}$ then we infer that either the null hypothesis does not describe the data, or that a rare (less than 5 percent probability) event has been observed.

The usual way to estimate a desired quantile, Q^θ for $0 < \theta < 1$, is to use the corresponding quantile of the sample. If $Y_1, \ldots, Y_n$ are independently sampled

from F, then we first sort them. Denote the sorted values, known as the ***order statistics*** by $Y_{(1)} \leqslant Y_{(2)} \leqslant \cdots \leqslant Y_{(n)}$ where (r) is the index of the r'th largest sample value. Then a simple quantile estimate is

$$\widetilde{Q}^\theta = Y_{(\lceil n\theta \rceil)}. \tag{2.24}$$

Sometimes we prefer a method that interpolates between consecutive order statistics. Several proposals of this type have been made in references given on page 36 of the chapter end notes. One such proposal is to take

$$\widehat{Q}^\theta = (1-\gamma)Y_{(t)} + \gamma Y_{(t+1)} \tag{2.25}$$

where

$$t = 1 + \lfloor (n-1)\theta \rfloor, \quad \text{and} \quad \gamma = (n-1)\theta - \lfloor (n-1)\theta \rfloor.$$

Equation (2.25) is meant for $0 < \theta < 1$ but it is well defined for $\theta = 0$ too. It is not well defined for $\theta = 1$, but $\lim_{\theta \to 1^+} \widehat{Q}^\theta = Y_{(n)}$, and so we may interpret $\widehat{Q}^1$ as $Y_{(n)}$.

It is intuitively clear that we should take n much larger than $1/\min(\theta, 1-\theta)$. Otherwise $\widehat{Q}^\theta$ will effectively be either $Y_{(1)}$ or $Y_{(n)}$ and will not be accurately estimated.

Example 2.2 (50 year wind)**.** As a purely synthetic example, consider a building that is located in a region where the strongest wind W in 50 years has an extreme value (see Example 4.9 in §4.2) distribution:

$$\mathbb{P}(W \leqslant w) = \exp(-\exp((w-52)/4)), \quad 0 < w < \infty.$$

Suppose that the load (pressure) this wind puts on the building is $L = CW^2$. The value of C depends on some properties of the building, the density of air, direction of the wind relative to the building geometry, positions of nearby buildings and other factors. To illustrate quantiles, we'll simply take $\log(C) \sim \mathcal{N}(-6, 0.05^2)$, independent of W. Figure 2.2 shows a histogram of $n = 10{,}000$ simulated values. The distribution has a heavy right tail. Even the logarithm of L (not shown) has positive skewness. The estimated 99'th percentile of L is near 12.37. A 99% confidence interval (derived below) for the 99'th percentile of the 50 year maximum wind speed goes from 12.04 to 12.70. This example is simple enough that even $n = 10^6$ realizations can be done quickly. The larger sample gives $\widehat{Q}^{0.99} \doteq 12.43$ with a 99% confidence interval $[12.39, 12.47]$. At the larger sample size, the reference lines for the 99% confidence interval almost overlap.

We may also want a confidence interval for an estimated quantile. Sample quantiles are not generally unbiased. Also, their variance depends on the probability density function of Y in a neighborhood of the true quantile value. Fortunately, we can get a confidence interval for Q^θ without having to work with either the mean or the variance of $\widehat{Q}^\theta$.

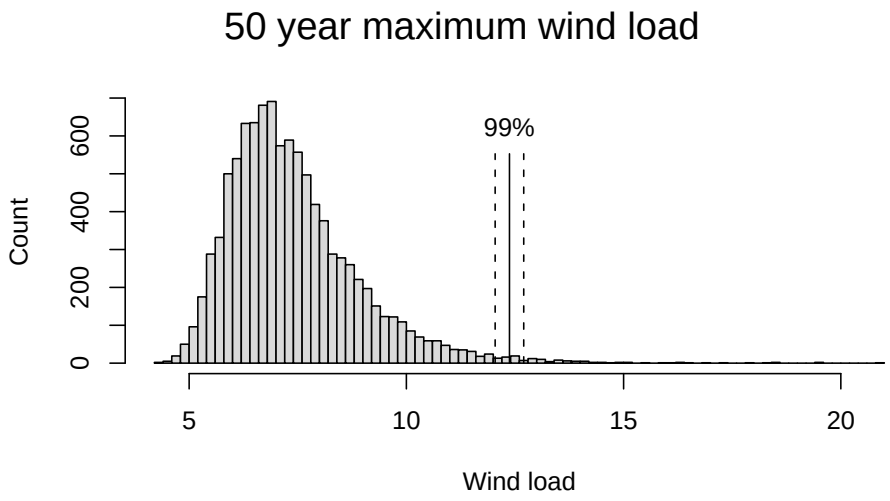

Figure 2.2: This figure shows the histogram of 10,000 simulated 50 year maximum wind forces. Reference lines show the 99'th percentile and a 99% confidence interval for that percentile.

First, suppose that we thought the 99'th percentile of L from Example 2.2 might be 11. That value is not very compatible with the histogram in Figure 2.2 because only 9727 of the values (or 97.27%) were below 11 and we expect 9900 below the quantile. If 11 were the true $Q^{0.99}$, then the number of simulated values below 11 would have the $\mathrm{Bin}(10000, 0.99)$ distribution. Because $\mathbb{P}(\mathrm{Bin}(10000, 0.99) \leqslant 9727) < 10^{-46}$ we can be very confident that $Q^{0.99} > 11$. We apply this idea to get confidence intervals for quantiles in general.

Let Y have a continuous distribution. Then $\mathbb{P}(Y \leqslant Q^\theta) = \theta$ holds and so $Q^\theta = \eta$ if and only if $\mathbb{E}(\mathbb{1}_{Y\leqslant\eta}) = \theta$. If θ is not inside the confidence interval for $\mathbb{E}(\mathbb{1}_{Y\leqslant\eta})$ then we can reject $Q^\theta = \eta$. As a result, we can obtain confidence intervals for Q^θ via confidence intervals for a binomial proportion. Of the methods from §2.4 we will work with the exact confidence intervals.

As a candidate value η increases, the number of Y_i below it only changes when η crosses one of the order statistics $Y_{(i)}$ of the sample. As a result, our confidence interval will take the form $[Y_{(L)}, Y_{(R)}]$ for integers L and R. If Y has a continuous distribution, then so does $Y_{(R)}$ and then

$$\mathbb{P}(Y_{(L)} \leqslant Q^\theta \leqslant Y_{(R)}) = \mathbb{P}(Y_{(L)} \leqslant Q^\theta < Y_{(R)}) = \mathbb{P}(L \leqslant X < R)$$

where $X = \sum_{i=1}^n \mathbb{1}_{Y_i\leqslant Q^\theta} \sim \mathrm{Bin}(n, \theta)$.

Now suppose that we want an interval with confidence $1-\alpha$ (i.e., $100(1-\alpha)\%$)

Algorithm 2.1 Confidence interval for a quantile.

given ordered sample $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ from a continuous dist'n, confidence level $1-\alpha$, quantile level $\theta \in (0,1)$, CDF $B_{n,\theta}$ of the $\mathrm{Bin}(n,\theta)$ distribution.

$L \leftarrow B^{-1}_{n,\theta}(\alpha/2)$
if $B_{n,\theta}(L) \leqslant \alpha/2$ **then**
  $L \leftarrow L+1$
$R \leftarrow B^{-1}_{n,\theta}(1-\alpha/2)+1$
deliver $[Y_{(L)}, Y_{(R)}]$

Cases $L=0$ and $R=n+1$ may arise when n is too small. We may then deliver $Y_{(0)}=-\infty$ and $Y_{(n+1)}=\infty$, respectively. For numerical reasons the test on L is not $B_{n,\theta}(L)=\alpha/2$, though $B_{n,\theta}(L)<\alpha/2$ is not possible mathematically.

where α is a small value like 0.01. Then we pick

$$\begin{aligned} L &= \max\left\{\ell \in \{0,1,\dots,n+1\} \mid \sum_{x=0}^{\ell-1}\binom{n}{x}\theta^x(1-\theta)^{n-x} \leqslant \alpha/2\right\}, \quad \text{and,} \\ R &= \min\left\{r \in \{0,1,\dots,n+1\} \mid \sum_{x=r}^{n}\binom{n}{x}\theta^x(1-\theta)^{n-x} \leqslant \alpha/2\right\}. \end{aligned} \tag{2.26}$$

Now $\mathbb{P}(L \leqslant X < R) \geqslant 1-\alpha/2-\alpha/2 = 1-\alpha$ and our confidence interval is $[Y_{(L)}, Y_{(R)}]$. Notice that we can find L and R before doing the simulation, and those values work for any continuous random variable Y. Equation (2.26) can give $L=0$ or $R=n+1$. These boundary cases are interpreted below.

Because we only have a discrete menu of choices for L and R, only a finite number of confidence levels $1-\alpha$ can be exactly attained given n and θ. The confidence intervals we obtain ordinarily have more than $1-\alpha$ probability of covering Q^θ. For the wind example, $L=9873$ and $R=9926$ so that

$$\begin{aligned} \mathbb{P}(L \leqslant X < R) &= 1-\mathbb{P}(X \leqslant 9872)-\mathbb{P}(X \geqslant 9926) \\ &\doteq 1-.0038-.0038 = 0.9924. \end{aligned}$$

For larger n, the attained confidence levels tend to get closer to the desired ones. With $n=10^6$ the attained confidence for the 99% interval around the 99'th percentile is about 99.006%.

The sums inside (2.26) are tail probabilities of the $\mathrm{Bin}(n,\theta)$ distribution. Let $B_{n,\theta}(x) = \mathbb{P}(\mathrm{Bin}(n,\theta) \leqslant x)$ be the cumulative distribution function (CDF) of X. The inverse of a CDF is defined in §4.1. Many computing environments include the inverse of the binomial CDF, and we can use it to get L and R without having to search for the solutions of (2.26). The result is presented in Algorithm 2.1 and proved in Exercise 2.13.

There are two special cases that could cause trouble. It is possible to get $L=0$. This happens when $\mathbb{P}(X=0) = (1-\theta)^n > \alpha/2$. In this case, we

have chosen n too small for the given α and θ. For $\ell = 0$, the tail probability in (2.26) sums over zero terms and hence is always below $\alpha/2$. We can interpret $Y_{(0)}$ as the smallest possible value for Y, which would be 0 in the wind example, but $-\infty$ if we had no prior lower bound on Y. At the other end, R as defined in (2.26) might be the artificially added value $n+1$. Once again this means that n is too small. We should take the upper confidence limit to be the largest possible value of Y, called $Y_{(n+1)}$ for convenience here. That value might be ∞ in the wind example.

The value $\alpha/2$ in (2.26) can be replaced by two values $\alpha_L \geqslant 0$ and $\alpha_R \geqslant 0$, for L and R respectively, with $\alpha_L + \alpha_R = \alpha$. For instance taking $\alpha_L = 0$ and $\alpha_R = \alpha$ yields a one-sided confidence interval $(-\infty, Y_{(R)}]$ for Q^θ.

2.6 Random sample size

We often find that the sample size in simple Monte Carlo is itself random. Suppose for example that we sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim p$ independently. In addition to studying $\mathbb{E}(f(\boldsymbol{X}))$ we might have a special interest in $\boldsymbol{X}$ that satisfy further conditions. For instance in Exercise 2.15 and Example 2.5 below, we investigate a model for the length of time that comets spend in the solar system. We might want to single out the longest lasting comets for further study. The number of long lasting comets that we get in our sample of n will be random.

From our n observations, suppose that we want to focus on only those that satisfy $\boldsymbol{X}_i \in A$ for some set A. The number of such observations that we get is

$$n_A = \sum_{i=1}^{n} A_i, \quad \text{where} \quad A_i = \begin{cases} 1, & \boldsymbol{X}_i \in A \\ 0, & \text{else.} \end{cases}$$

Suppose that we want to study $\mu_A = \mathbb{E}(f(\boldsymbol{X}) \mid \boldsymbol{X} \in A)$. The simple and direct approach is to proceed as if we had intentionally obtained n_A observations from the distribution of $\boldsymbol{X}$ given that $\boldsymbol{X} = A$, that is, as n_A observations from the density $p_A(\boldsymbol{x}) = p(\boldsymbol{x})\mathbb{1}_{\boldsymbol{x}\in A}/\int_A p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. We estimate μ_A by

$$\hat{\mu}_A = \frac{1}{n_A}\sum_{i=1}^{n} A_i Y_i$$

and setting

$$s_A^2 = \frac{1}{n_A - 1}\sum_{i=1}^{n} A_i (Y_i - \hat{\mu}_A)^2$$

our 99% confidence interval is $\hat{\theta} \pm 2.58 s_A/\sqrt{n_A}$. We do have to assume that n_A is large enough for $\hat{\mu}_A$ and s_A to be reasonable estimates, but we do not need to take account of the fact that n_A was random and might have been different than the value we got. This intuitively reasonable answer is almost identical to what we get from a ratio estimation derivation in §2.7.

2.7 Estimating ratios

We often want to estimate a ratio $\theta = \mathbb{E}(Y)/\mathbb{E}(X)$ for some jointly distributed random variables X and Y. Sometimes our interest is in the ratio itself. In other settings we may know $\mathbb{E}(X)$ and then an estimate of $\mathbb{E}(Y)/\mathbb{E}(X)$ can lead to an efficient way to estimate $\mathbb{E}(Y)$. See the discussion of the ratio estimator in §8.9.

Weighted sample means often give rise to ratio estimation problems. In §9.2 on self-normalized importance sampling we write $\mathbb{E}_p(f(X)) = \mathbb{E}_q(w(X)f(X))/\mathbb{E}_q(w(X))$ for a weighting function $w(X) \geqslant 0$ where $\mathbb{E}_p$ is expectation for $X \sim p$ and $\mathbb{E}_q$ is expectation for $X \sim q$. The ratio estimator has wf playing the role of Y and w the role of X, both under sampling from q. This ratio estimator may be more accurate than an ordinary estimator sampled from p. The special case of binary weights is considered below in Example 2.3. The ratio estimate of the weighted mean will give us the same answer if we replace $w(X)$ by $\widetilde{w}(X) = cw(X)$ for any $c > 0$. That estimate is useful in settings where we can compute $\widetilde{w}(X)$ but not $w(X)$, because the constant of proportionality, c, is unknown.

The natural way to estimate θ is to sample n independent pairs (X_i, Y_i) from the target distribution and then take

$$\hat{\theta} = \bar{Y}/\bar{X} \tag{2.27}$$

where $\bar{X} = (1/n)\sum_{i=1}^n X_i$ and $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$.

To use simple Monte Carlo here, we face two problems. First $\mathbb{E}(\hat{\theta}) \neq \theta$ in general. Ordinarily, this bias becomes unimportant for large n. However that will still leave us with the task of estimating the variance of $\hat{\theta}$ to form a confidence interval for θ.

We can approximate the mean and variance of $\hat{\theta}$ using the delta method. For background on the delta method in general, see the discussion on page 36 of the chapter end notes. For now, we note that we seek a confidence interval for $f(\mathbb{E}(X), \mathbb{E}(Y))$ centered on $f(\bar{X}, \bar{Y})$, where $f(x, y) = y/x$. The delta method solves this problem for general smooth functions f of one, two, or more means, based on Taylor expansions of f.

Applying (2.38) from the end notes to $f(\bar{X}, \bar{Y})$ we find the following approximation to $\mathrm{Var}(\hat{\theta})$:

$$\frac{1}{n}\Big(\sigma_x^2 f_x^2 + \sigma_y^2 f_y^2 + 2\rho_{xy}\sigma_x\sigma_y f_x f_y\Big), \tag{2.28}$$

where $f_x = (\partial f/\partial x)(\mu_x, \mu_y)$, and $f_y = (\partial f/\partial y)(\mu_x, \mu_y)$. The quantities σ_x^2, σ_y^2 and ρ_{xy} are the variances of X and Y and the correlation of X with Y, respectively.

For the ratio estimator, equation (2.28) simplifies to

$$\frac{1}{n}\frac{\mathbb{E}((Y-\theta X)^2}{\mu_x^2}. \tag{2.29}$$

See Exercise 2.9. After plugging in estimates of μ_x and the other unknowns, the delta method leads to the variance estimate

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{n^2\bar{X}^2}\sum_{i=1}^{n}(Y_i - \hat{\theta}X_i)^2. \tag{2.30}$$

We could equally well have arrived at $n(n-1)$ in the denominator instead of n^2 by making our approximations in a slightly different way, but this distinction makes no important difference for large n. The 99% confidence interval for the ratio θ, from the delta method, is

$$\hat{\theta} \pm 2.58\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}. \tag{2.31}$$

From equation (2.44) of the end notes, we get the following approximation to the bias of $\hat{\theta}$:

$$\mathbb{E}(\hat{\theta} - \theta) \doteq \frac{1}{2n}\bigl(f_{xx}\sigma_x^2 + f_{yy}\sigma_y^2 + 2f_{xy}\rho\sigma_x\sigma_y\bigr) \tag{2.32}$$

where f_{xx}, f_{xy}, and f_{yy} are the obvious second order partial derivatives of f, evaluated at (μ_x, μ_y). For $f(x, y) = y/x$ equation (2.32) simplifies to

$$\mathbb{E}(\hat{\theta} - \theta) \doteq \frac{1}{n\mu_x^2}\bigl(\theta\sigma_x^2 - \rho\sigma_x\sigma_y\bigr). \tag{2.33}$$

Because the bias is $O(1/n)$ while the root mean squared error is of order $1/\sqrt{n}$ the confidence interval for θ ignores the bias.

An alternative approach to ratio estimation starts by writing $\theta = \mathbb{E}(Y)/\mathbb{E}(X)$ as $\mathbb{E}(Y - \theta X) = 0$. For any candidate value θ we can make a confidence interval I_θ for $\mathbb{E}(Y - \theta X)$. Then the confidence interval for θ is $\{\theta \mid 0 \in I_\theta\}$. This is the Fieller solution (Fieller, 1954) and it has the same logic as the method in §2.5 where we constructed confidence intervals for a quantile using those for a proportion. Most Monte Carlo analysis of ratio estimates uses the simpler delta method interval (2.31) based on the variance estimate (2.30).

Example 2.3 (Conditional expectation)**.** Consider a special form of the weight function

$$w(X) = \begin{cases} 1, & X \in A \\ 0, & X \notin A \end{cases}$$

where A is a set with $\mathbb{P}(X \in A) > 0$. Then $\theta = \mathbb{E}(f(X)w(X))/\mathbb{E}(w(X))$ is just $\hat{\mu}_A = \mathbb{E}(f(X) \mid X \in A)$ from §2.6. Now suppose that we have sampled n points $X_1, \ldots, X_n$ in a Monte Carlo calculation. Introduce $A_i = w(X_i)$ and $Y_i = f(X_i)$ as shorthand. In this Monte Carlo, the event A happened $n_A = \sum_{i=1}^n A_i$ times. When $n_A \geqslant 1$, we may estimate the expected value of Y given that event A has occurred via the ratio estimate

$$\hat{\theta} = \frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i}. \tag{2.34}$$

The ratio estimate $\hat{\theta}$ matches the estimate $\hat{\mu}_A$ that we had in the random sample size context of §2.6. The variance estimate is virtually identical too. All that changes is that the variance estimate (2.30) leads us to divide the sum of squares by n_A instead of $n_A - 1$. See Exercise 2.10. That is, the ratio estimation 99% confidence interval is $\hat{\theta} \pm 2.58\hat{\sigma}_A/\sqrt{n_A}$ where $\hat{\sigma}_A^2 = (n_A - 1)s_A^2/n_A$.

2.8 When Monte Carlo fails

Monte Carlo methods are extremely robust compared to alternatives, but there are still ways in which they can go wrong. One problem, alluded to in §2.1, is that μ might not even exist. For $\mu = \mathbb{E}(Y)$ to exist, and be finite, we must have $\mathbb{E}(|Y|) < \infty$.

When $\mathbb{E}(|Y|) = \infty$ then it is possible that $\mathbb{E}(Y) = +\infty$ or $\mathbb{E}(Y) = -\infty$ or that $\mathbb{E}(Y)$ is not even defined as a member of $[-\infty, \infty]$. The latter case arises when $\mathbb{E}(\max(Y, 0)) = \mathbb{E}(\max(-Y, 0)) = \infty$.

Example 2.4 (St. Petersburg paradox)**.** We can illustrate $\mu = \infty$ with the St. Petersburg paradox. Suppose that a gambler is offered the following proposition. A fair coin will be tossed until heads comes up for the first time. Let X be the number of tosses made. If $X = x$, then the gambler will get $\$2^x$. For independent coin tosses $p_k = \mathbb{P}(X = k) = 2^{-k}$ for $k \geqslant 1$ and the expected payoff in dollars is $\mu = \sum_{k=1}^{\infty} p_k 2^k = \sum_{k=1}^{\infty} 2^{-k} 2^k = \infty$. The paradox is not the mere fact that $\mu = \infty$. The paradox is that there is no good answer to the question of how much a gambler should be willing pay for a chance to play the game. Any finite entry price, no matter how large, leaves the gambler with a positive expected return, but that does not mean it is wise to pay a very large sum to play this game.

Figure 2.3 plots the value of $\hat{\mu}_n$ versus n for a simulation of 10,000 draws from the St. Petersburg paradox problem. Every once in a while there is a sharp upward jump in $\hat{\mu}_n$. Between jumps $\hat{\mu}_n$ drifts down towards zero. For y_n to be large enough to cause a jump, it must be of comparable size to $y_1 + \ldots + y_{n-1}$. Such jumps have to keep coming, because $\hat{\mu}_n$ is going to ∞.

In a problem with $\mu = \infty$, we have $\mathbb{P}(\hat{\mu}_n \to \infty) = 1$ by the law of large numbers. But when all of the x_i are always finite, then $\mathbb{P}(\hat{\mu}_n = \infty) = 0$ for all n. That is, we can't just wait until $\hat{\mu}_n = \infty$ and then declare that $\mu = \infty$. Furthermore, the central limit theorem does not apply at all here. The interval $\hat{\mu}_n \pm \Phi^{-1}(0.995)s/\sqrt{n}$ has probability 0 of containing μ for any n.

For the St. Petersburg paradox it was pretty easy to know from the beginning that μ was infinite. In other settings μ might be infinite without our knowing it. A common way that infinite means might arise is in ratios of random variables: $\mu = \mathbb{E}(Y/X)$ might not exist because Y/X can become large for small $|X|$ not just large Y.

Ratios very often arise as derived quantities and in importance sampling (Chapter 9). The quantity $\theta = \mathbb{E}(Y)/\mathbb{E}(X)$ can differ by an arbitrarily large

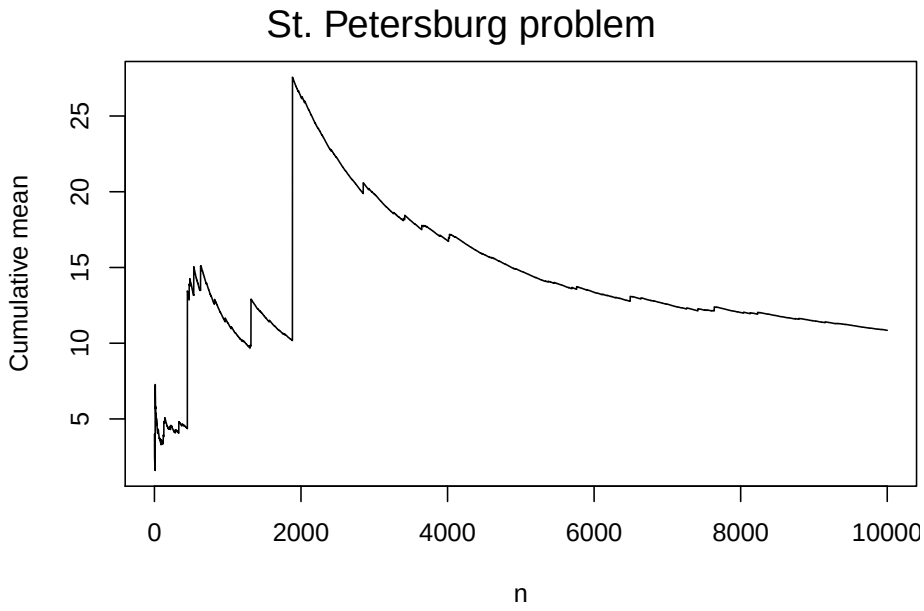

Figure 2.3: This figure shows the running mean of 10,000 simulated draws from the distribution in the St. Petersburg paradox.

amount from $\mathbb{E}(Y/X)$. But when we want to estimate such a ratio, we ordinarily use a ratio $\hat{\theta} = \bar{Y}/\bar{X}$ of sample averages.

In ratios, for μ to be finite we ordinarily need the mean of the numerator to be finite and the denominator should not have a positive density at 0. Small changes in the distribution of the denominator can turn a problem with finite expected ratio into one with infinite expected ratio.

Example 2.5 (Long lived comets)**.** Hammersley and Handscomb (1964) present a Monte Carlo calculation for the lifetime of a long lived comet. The comet has an energy level x, defined in such a way that positive values give elliptical orbits, and negative values give hyberbolic orbits. A comet with energy $x < 0$ takes time $(-x)^{-3/2}$ to complete one orbit, while one with $x > 0$ just leaves the solar system. A small portion of a comet's orbit takes place near the planets orbiting the sun. Gravitational interactions with planets change the energy level of the comet. Their model is that x changes to $x+Z$ where $Z \sim \mathcal{N}(0, \sigma^2)$. By choosing the units of energy appropriately, they can use $\sigma = 1$. A comet that starts with energy x_0 will stay in the solar system for time

$$T = \sum_{j=0}^{m-1} (-x_j)^{-3/2}$$

where $x_{j+1} = x_j + z_j$ and $m = \min\{\, j \mid x_j > 0\}$ indexes the first orbit to obtain a positive energy. Because the number m of orbits to count is itself random,

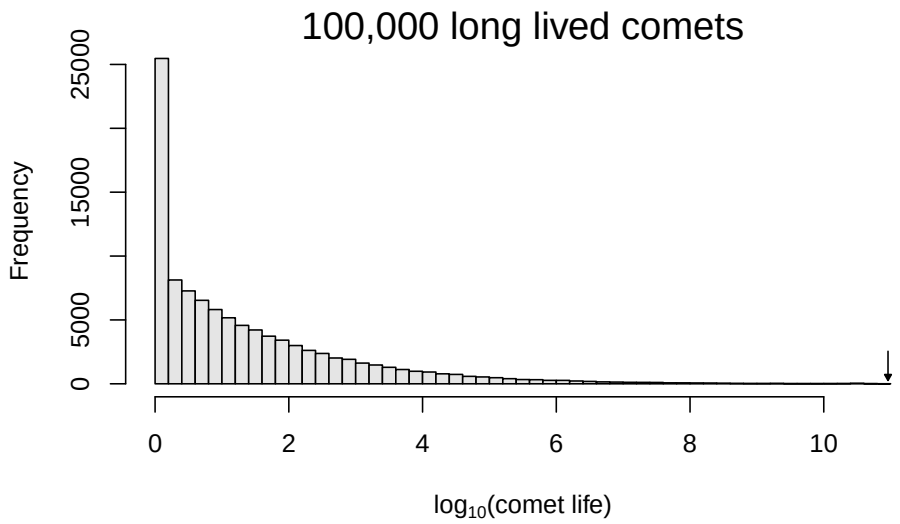

Figure 2.4: This figure shows the orbital lives of 100,000 comets simulated as in Example 2.5, each starting with initial energy $x_0 = -1$. A reference arrow marks the longest observed comet life. Each comet was followed for at most 10^6 orbits. There were 100 comets that survived past 10^6 orbits.

it is hard to study the distribution of T analytically. An asymptotic analysis in Hammersley (1961) shows that $\mathbb{P}(T > t)$ is approximately proportional to $t^{-2/3}$ for very large t. As a result, T has a probability density function that decays like $t^{-5/3}$. Therefore $\mathbb{E}(T) = \mathbb{E}(T^2) = \infty$, and so T does not have finite variance. Figure 2.4 shows some sampled comet lives. The quantity T is well suited to Monte Carlo sampling. We can easily get estimates and confidence intervals for $\mathbb{P}(T > t_0)$.

A second kind of failure for Monte Carlo comes when the mean is finite, but the variance is infinite. Then the law of large numbers still assures convergence to the right answer, but we lose the $O(\sqrt{n})$ rate as well as the confidence intervals from the central limit theorem.

If $\mathbb{E}(|Y|) < \infty$ and $\mathrm{Var}(Y) = \infty$ then it is still possible to estimate $\mu = \mathbb{E}(Y)$ and get a confidence interval for it. But to do so requires very specialized methods outside the scope of this book. See Peng (2004). With Monte Carlo methods we have many ways to reformulate the problem, preserving the finite expectation while obtaining a finite variance. Importance sampling, in Chapter 9 is one such method.

We can get a hint that $\mathbb{E}(|Y|) = \infty$ or that $\sigma^2 = \infty$ from plots like Figure 2.3. But Monte Carlo is not the right tool for determining whether these moments exist. Only mathematical analysis can be conclusive. The plot could be wrong by suggesting $\mu = \infty$ when Y has a long tail that ends at some very large value. Conversely, our Monte Carlo sample could give every indication that μ is well

behaved but μ could fail to exist because some small region of space, with say probability 10^{-23} has not yet been sampled. Problems of that sort can arise when importance sampling is poorly applied.

Even when we observe $y_i = \infty$ for some i we can still not be sure that $\mu = \infty$. The problem is that computer arithmetic is discrete. Suppose for example that $Y = |X - 1/2|^{-1/5}$ for $X \sim \mathbf{U}(0,1)$. If $X = 1/2$ is one of the possible values from our random number generator we could get $\hat{\mu}_n = \infty$, even though we can prove mathematically that μ is finite.

In other applications σ^2 exists but $\mathbb{E}(|Y|^3)$ or $\mathbb{E}(|Y|^4)$ is infinite. These are comparatively mild problems. The central limit theorem still holds but it sets in more slowly, as the speed is governed by $\mathbb{E}(|Y-\mu|^3)/\sigma^3$. If $\sigma < \infty$, then s^2 approaches σ^2 by the law of large numbers. But for s^2 to have an RMSE of $O(n^{-1/2})$ requires $\mathbb{E}(|Y|^4) < \infty$.

2.9 Chebychev and Hoeffding intervals

Sometimes we want to estimate $\mu = \mathbb{E}(Y)$ and we have extra knowledge about Y, such as a known upper bound for $\sigma^2 = \mathrm{Var}(Y)$. This knowledge allows us to determine n before doing any sampling.

Suppose that $\mathrm{Var}(Y) = \sigma_0^2$ is known. Then $\mathrm{Var}(\hat{\mu}) = \sigma_0^2/n$. It follows from Chebychev's inequality that

$$\mathbb{P}\big(|\hat{\mu} - \mu| \geqslant k\sigma_0\sqrt{n}\,\big) \leqslant \frac{1}{k^2}. \tag{2.35}$$

Then we can get a conservative 99% confidence interval for μ by taking

$$\hat{\mu} \pm \frac{10\sigma_0}{\sqrt{n}}. \tag{2.36}$$

So long as $\mathrm{Var}(Y) = \sigma^2 \leqslant \sigma_0^2$, the interval (2.36) has at least 99% probability of containing μ.

The width of this interval is $20\sigma_0/\sqrt{n}$. If we want a 99% confidence interval of a given width ε then we can get it by sampling $n = \lceil 400\sigma_0^2/\varepsilon^2 \rceil$ values of Y_i.

The Chebychev interval is not useful if we don't know σ. If we replace σ_0 by an estimate s, then equation (2.35) need not hold. We lose the gaurantee.

There is another difficulty with the Chebychev interval. The gauranteed coverage comes at a high cost. For 99% confidence we really only need to multiply $\sigma/\sqrt{n}$ by 2.58 instead of 10. From the central limit theorem the interval (2.36) has coverage that tends to $1 - 2\Phi(-10) \doteq 1 - 1.52\times10^{-23}$, much better than the 99% we claim. We have used $(10/2.58)^2 \doteq 15.0$ times as much computation as we needed to for 99% confidence.

A second kind of extra knowledge takes form of known bounds $a_i \leqslant Y_i \leqslant b_i$ for $i = 1, \ldots, n$. For IID Y_i we have common values $a_i = a$ and $b_i = b$, but the more general case is not much more complicated.

Theorem 2.2 (Hoeffding's inequality)**.** *Let $Y_1, \dots, Y_n$ be independent random variables with $\mathbb{P}(a_i \leqslant Y_i \leqslant b_i) = 1$ for $i = 1, \dots, n$. Then for $\epsilon > 0$*

$$\mathbb{P}\left(\sum_i (Y_i - \mathbb{E}(Y_i)) \geqslant \epsilon\right) \leqslant e^{-2\epsilon^2/\sum_{i=1}^n (b_i - a_i)^2}, \quad \text{and}$$

$$\mathbb{P}\left(\sum_i (Y_i - \mathbb{E}(Y_i)) \leqslant \epsilon\right) \leqslant e^{-2\epsilon^2/\sum_{i=1}^n (b_i - a_i)^2}.$$

Proof. This statement is from Devroye et al. (1996). The original proof is in Hoeffding (1963). □

Corollary 2.1. *Let $Y_1, \dots, Y_n$ be independent random variables with mean μ such that $a \leqslant Y_i \leqslant b$ for finite a and b. Let $\hat{\mu} = (1/n)\sum_{i=1}^n Y_i$ and $\delta \in (0, 1)$. Then for $\varepsilon > 0$,*

$$\mathbb{P}(|\hat{\mu} - \mu| \geqslant \varepsilon/2) \leqslant \delta$$

when $n \geqslant 2(b-a)^2 \log(2/\delta)/\varepsilon^2$.

Proof. The event $|\hat{\mu} - \mu| \geqslant \varepsilon/2$ is the same as $|\sum_{i=1}^n (Y_i - \mu)| \geqslant \epsilon \equiv n\varepsilon/2$. From Theorem 2.2, $\mathbb{P}(|\hat{\mu} - \mu| \geqslant \varepsilon/2) \leqslant 2\exp(-(1/2)\varepsilon^2 n/(b-a)^2)$. For n of the stated size this probability is no more than $2\exp(-\log(2/\delta)) = \delta$. □

Corollary 2.1 shows how large n must be, for the interval $\hat{\mu} \pm \varepsilon/2$ to have a gauranteed confidence level of $100(1-\delta)\%$ or more. For 99% confidence we take $\delta = 0.01$. Then we need $n \geqslant 2\log(2/\delta)(b-a)^2/\varepsilon^2 \doteq 10.6(b-a)^2/\varepsilon^2$.

Chapter end notes

Order symbols O, o, O_p, o_p

The root mean square error in Monte Carlo is $\sigma/\sqrt{n}$ when n function evaluations are used. By contrast, Simpson's rule §7.2 for integration of a function f with continuous fourth derivative $f^{(4)}$ on $[0, 1]$ has a deterministic error equal to $-f^{(4)}(z)/(180n^4)$ for some z with $0 < z < 1$.

For two functions f and g, we write $f(n) = O(g(n))$ as $n \to \infty$ if there are constants C and n_0 such that $|f(n)| \leqslant Cg(n)$ whenever $n \geqslant n_0$. Therefore the RMSE in Monte Carlo is $O(n^{-1/2})$ as $n \to \infty$ whenever $\sigma < \infty$ and the error in a one dimensional Simpson's rule is $O(n^{-4})$ as $n \to \infty$ whenever f is smooth enough.

Big O notation allows us to focus on the rates of convergence as $n \to \infty$. It hides the constant C as well as the threshold n_0. We do not always know their values even when we know they exist. We often drop the clause 'as $n \to \infty$' when it is understood. The iterated Simpson's rule on $[0, 1]^d$ has error $O(n^{-4/d})$, again for smooth enough f.

The limiting operation does not have to be as $n \to \infty$ through integers. We can have $f(x) = O(g(x))$ as $x \to 0$ or as $x \to \infty$. In the former case, the limit might be $x \to 0+$ meaning a limit from the right. There must still be a

bounding constant and a threshold for the limit we are taking. For example $f(x) = O(g(x))$ as $x \to 0+$ means that for some $\epsilon > 0$ and $C < \infty$ we have $|f(x)| \leqslant Cg(x)$ whenever $0 \leqslant x \leqslant \epsilon$.

Given a choice of two competing methods, we ordinarily favor the one with the better error rate. But a better big O rate does not imply a better result. Suppose that two methods have errors $E_n = O(n^{-r})$ and $F_n = O(n^{-s})$ with $r > s > 0$. Then for $n \geqslant n_0$ (the larger of the two thresholds) we have $|E_n| \leqslant Cn^{-r}$ and $|F_n| \leqslant Dn^{-s}$. For a given integer n, we might guess that $|E_n| < |F_n|$, but it does not have to be true. Three things can go wrong. First, n might not be larger than the threshold n_0. Second, the implied constant C might be much larger than D, so that $Cn^{-r} > Dn^{-s}$. Finally, the convergence rate for $|F_n|$ could be much better than n^{-s}. For example, when $E_n = n^{-s-2}$ we still have $E_n = O(n^{-s})$.

Small o notation is a related concept. For two functions f and g, we write $f(n) = o(g(n))$ as $n \to \infty$ if $\lim_{n\to\infty} f(n)/g(n) = 0$. While O means 'asymptotically no larger than', o means 'asymptotically smaller than'.

When $f(n) \in \mathbb{R}^s$ for $s > 1$ then $f(n) = O(g(n))$ means that $\|f(n)\| = O(g(n))$. Similarly $f(n) = o(g(n))$ means that $\|f(n)\| = o(g(n))$.

Some random quantities are of typical size $n^{-1/2}$ but are unbounded and hence not $O(n^{-1/2})$. For example $n^{-1/2}$ times a $\mathcal{N}(0,1)$ random variable fits this description. For a sequence of random variables X_n with $n \geqslant 1$ we write $X_n = O_p(c_n)$ as $n \to \infty$, if for any $\epsilon > 0$ there is $B = B_\epsilon < \infty$ such that

$$\limsup_{n\to\infty} \mathbb{P}(|X_n| > Bc_n) < \epsilon.$$

For example when $X_n = (1/n)\sum_{i=1}^n Y_i$ for IID Y_i with mean μ and variance $\sigma^2 < \infty$, then $X_n - \mu = O_p(n^{-1/2})$. Similarly, we write $X_n = o_p(c_n)$ if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon c_n) = 0.$$

Variance updating formulas

Chan et al. (1983) provide an analysis of the numerical stability of variance updates like (2.18) which go back to at least Welford (1962). Let ε be the machine precision. This is the smallest $x > 0$ such that when $1 + x$ is computed it comes out larger than 1. For floating point arithmetic in double precision, ε is about 2×10^{-16} while for single precision (not recommended) it is about 10^{-7}. They note that the relative error in computing s^2 by updating is at most $n\varepsilon\sqrt{1 + \hat\mu^2 n/S} = n\varepsilon\sqrt{1 + \hat\mu^2/\hat\sigma^2}$.

Single precision arithmetic is seen to be pretty bad for this problem because n might not be small compared to $1/\varepsilon$.

If extreme care needs to be taken in computing the variance because n is not small compared to the inverse of the double precision ε then there are more robust numerical methods than the update (2.18). These methods are based on summation techniques that sum data by first summing pairs of values, then pairs of pairs and so on. The space required for the bookkeeping to maintain

all of the partially summed data grows as $\log_2(n)$. Details are in Chan et al. (1983).

Empirical quantiles

Hyndman and Fan (1996) present nine different proposals for picking an empirical quantile. Equation (2.24) is a simple inverse of the empirical CDF, their method 1. The estimate in (2.25) is their method 7, which originated with Gumbel (1939) and is the default in R. They discuss various desirable properties that an empirical quantile should have, and give the history.

For further discussion of confidence intervals for quantiles, see Hahn and Meeker (1991). They give more attention to smaller values of n and in some such cases one can get a meaningfully shorter interval by using unequal upper and lower tail probabilities α_1 and α_2 with $\alpha_1 + \alpha_2 = \alpha$.

The delta method

Here we look at the delta method non-rigorously. For a rigorous treatment see Lehmann and Romano (2005, Chapter 11). Let $\boldsymbol{X} = (X_1, \dots, X_d) \in \mathbb{R}^d$ be a random vector with distribution F, and let $\theta = f(\mathbb{E}(\boldsymbol{X})) \in \mathbb{R}$ for a function f. We assume that $\mu = \mathbb{E}(\boldsymbol{X}) \in \mathbb{R}^d$ and $\mathrm{Var}(\boldsymbol{X}) = \Sigma \in \mathbb{R}^{d\times d}$ are both finite. The variance of X_j is written σ_j^2 and the correlation of X_j and X_j is ρ_{jk}.

From a sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ of independent random vectors with distribution F, we form $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$. Then the usual estimate of $\theta = f(\mu)$ is $\hat\theta = f(\bar{\boldsymbol{X}})$. The ***delta method*** gives an approximate confidence interval for θ that becomes increasingly accurate as n increases. It is based on Taylor expansion and the central limit theorem.

The first order Taylor expansion of $f(\bar{\boldsymbol{X}})$ around μ is

$$f(\bar{\boldsymbol{X}}) \doteq f(\mu) + \sum_{j=1}^d (\bar{X}_j - \mu_j) f_j(\mu) \tag{2.37}$$

where $f_j = \partial f/\partial X_j$. In the delta method, we approximate $f(\bar{\boldsymbol{X}})$ by the right hand side of (2.37) and then approximate the distribution of that right hand side using the central limit theorem.

The right hand side of (2.37) has mean $f(\mu)$ and variance

$$\frac{1}{n}\biggl(\sum_{j=1}^d f_j(\mu)^2\sigma_j^2 + 2\sum_{j=1}^{d-1}\sum_{k=j+1}^d f_j(\mu)f_k(\mu)\rho_{jk}\sigma_j\sigma_k\biggr). \tag{2.38}$$

We can simplify equation (2.38). Let $\nabla f = (\nabla f)(\mu) = (f_1(\mu), \cdots, f_d(\mu))^{\mathsf{T}} \in \mathbb{R}^d$ be the gradient of f at μ, represented as a column vector. Then the variance of our approximation to $f(\bar{\boldsymbol{X}})$ is

$$\widetilde{\mathrm{Var}}(\hat\theta) = \frac{1}{n}(\nabla f)^{\mathsf{T}}\Sigma\,(\nabla f). \tag{2.39}$$

Approximation (2.39) is not usable, unless we know μ and Σ. Typically we use estimates $\hat{\mu} = \bar{\boldsymbol{X}}$ and $\widehat{\Sigma} = (n-1)^{-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \hat{\mu})(\boldsymbol{X}_i - \hat{\mu})^{\mathsf{T}}$ for these moments. Then putting $\widehat{\nabla} f = (\nabla f)(\hat{\mu})$ the delta method variance estimate is

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{n}(\widehat{\nabla} f)^{\mathsf{T}}\widehat{\Sigma}\,(\widehat{\nabla} f), \tag{2.40}$$

which we can compute, assuming that we can differentiate f at $\bar{\boldsymbol{X}}$.

From the central limit theorem, $\sqrt{n}(\hat{\theta} - \theta) \mathrel{\dot\sim} \mathcal{N}\big(0, (\nabla f)^{\mathsf{T}}\Sigma\,(\nabla f)\big)$. Substituting estimates of μ and Σ leads to

$$\sqrt{n}(\hat{\theta} - \theta) \mathrel{\dot\sim} \mathcal{N}\big(0, (\widehat{\nabla} f)^{\mathsf{T}}\widehat{\Sigma}\,(\widehat{\nabla} f)\big). \tag{2.41}$$

Replacing μ by $\hat{\mu}$ and Σ by $\hat{\Sigma}$ makes only a negligible difference, for large n. A formal justification of this substitution follows from Slutsky's theorem (Lehmann and Romano, 2005, Chapter 11) on page 38 of these end notes. The result is an approximate 99% confidence interval

$$\hat{\theta} \pm \frac{2.58}{\sqrt{n}}\sqrt{(\widehat{\nabla} f)^{\mathsf{T}}\widehat{\Sigma}\,(\widehat{\nabla} f)} \tag{2.42}$$

for θ.

Equation (2.42) is based on three approximations: the Taylor expansion (2.37), the central limit theorem, and the substitution of $(\hat{\mu}, \hat{\Sigma})$ for (μ, Σ). If f is differentiable at μ, then the errors in (2.37) are small compared to $\bar{\boldsymbol{X}} - \mu$, which in turn is usually very small when n is large, and as a result the Taylor expansion is good enough. If Σ exists, then the central limit theorem applies and furthermore, the sample moments $\hat{\mu}$ and $\hat{\Sigma}$ approach their population counterparts.

When $\nabla f(\mu) = 0$ then the approximation (2.41) yields $\sqrt{n}(\hat{\theta} - \theta) \mathrel{\dot\sim} \mathcal{N}(0, 0)$. In this circumstance $\mathbb{P}(\sqrt{n}|\hat{\theta} - \theta| > \epsilon) \to 0$ for any $\epsilon > 0$, but (2.41) fails to provide a confidence interval. If Σ is positive definite, and ∇f is nonzero, then we avoid this degenerate case and get a usable confidence interval.

The delta method can also be used to get approximations to the bias $\mathbb{E}(\hat{\theta} - \theta)$. The first order Taylor expansion (2.37) does not help in estimating the bias of $\hat{\theta}$. The expectation of the right hand side is $f(\mu)$. To study the bias we expand to second order,

$$f(\bar{\boldsymbol{X}}) \doteq f(\mu) + \sum_{j=1}^{d}(\bar{X}_j - \mu_j)f_j(\mu) + \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d}(\bar{X}_j - \mu_j)(\bar{X}_k - \mu_k)f_{jk}(\mu) \tag{2.43}$$

where $f_{jk} = \partial^2 f/\partial X_j \partial X_k$. Taking the expected value of the right hand side of (2.43) leads to the approximation

$$\mathbb{E}(\hat{\theta} - \theta) \doteq \frac{1}{2n}\sum_{j=1}^{d}\sum_{k=1}^{d} f_{jk}(\mu)\rho_{jk}\sigma_j\sigma_k. \tag{2.44}$$

The bias of $\hat{\theta}$ is typically $O(n^{-1})$. The mean squared error is $\mathbb{E}((\hat{\theta}-\theta)^2) = \mathrm{Var}(\hat{\theta}) + \big(\mathbb{E}(\hat{\theta}-\theta)\big)^2$. The variance term is $O(n^{-1})$ while the bias squared is $O(n^{-2})$. As a result, for large n, our error is usually dominated by the variance, not the bias.

In rare circumstances, we may know some of the moments of $\boldsymbol{X}$ that appear in equations (2.38) and (2.39). For the ratio estimator $\hat{\theta} = \bar{Y}/\bar{X}$ of §2.7, there has been a long debate on whether to use $(n\bar{X})^{-2}\sum_{i=1}^n(Y_i-\hat{\theta}X_i)^2$ or $(n\mu_x)^{-2}\sum_{i=1}^n(Y_i-\hat{\theta}X_i)^2$ in cases where μ_x is known. Neither estimator is always better than the other. An analysis in Qin and Li (2001), for survey sampling applications, shows that the standard approach using $\bar{X}$ is often better than using μ_x.

Slutsky's theorem

Let Y_n and Z_n be two sequences of random variables. When Z_n converges to a contant value τ in the way defined below, then we expect that $Y_n + Z_n$, $Y_n - Z_n$, Y_n/Z_n, Y_nZ_n and so on should behave like $Y_n+\tau$, $Y_n-\tau$, Y_n/τ, $Y_n\tau$ respectively for large n. Slutsky's theorem gives us conditions under which we can simply plug in an estimate Z_n of an unknown quantity τ without changing the asymptotic distribution.

For example, if a central limit theorem applies to $\sqrt{n}(\bar{Y}-\mu)/\sigma$ and we replace the unknown σ by an estimate $\hat{\sigma}$ we would like to get the same limiting normal distribution for $\sqrt{n}(\bar{Y}-\mu)/\hat{\sigma}$.

The sequence Y_n of random variables ***converges in distribution*** to the random variable Y, as $n\to\infty$, if $\mathbb{P}(Y_n \leqslant y) \to \mathbb{P}(Y \leqslant y)$ holds at every y where $F(y) \equiv \mathbb{P}(Y \leqslant y)$ is continuous. We write this as $Y_n \xrightarrow{\mathrm{d}} Y$. Note that Y_n does not have to be close to Y, it is only the distribution of Y_n that has to become close to the distribution of Y.

The sequence Z_n ***converges in probability*** to the value τ as $n\to\infty$, if $\mathbb{P}(|Z_n-\tau|>\epsilon)\to 0$ holds for all $\epsilon>0$. Then $Z_n \xrightarrow{\mathrm{d}} Z$ where Z is a degenerate random variable with $\mathbb{P}(Z=\tau)=1$. We write this convergence as $Z_n \xrightarrow{\mathrm{d}} \tau$.

Theorem 2.3 (Slutsky's Theorem)**.** *Suppose that random variables $Y_n \xrightarrow{\mathrm{d}} Y$ and $Z_n \xrightarrow{\mathrm{d}} \tau$. Then $Y_n+Z_n \xrightarrow{\mathrm{d}} Y$ and $Y_nZ_n \xrightarrow{\mathrm{d}} \tau Y$. If $\tau\neq 0$ then $Y_n/Z_n \xrightarrow{\mathrm{d}} Y/\tau$.*

Proof. Lehmann and Romano (2005, Chapter 11). □

Knight (2000, Chapter 3) gives a more general version where the conclusion is that $g(Y_n,Z_n) \xrightarrow{\mathrm{d}} g(Y_n,\tau)$ where g is a continuous function.

For the CLT with an estimated standard deviation, suppose that $0<\sigma<\infty$, as is usual. Then

$$\sqrt{n}\,\frac{\bar{Y}-\mu}{\sigma} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1) \quad \text{and} \quad \frac{s}{\sigma} \xrightarrow{\mathrm{d}} 1$$

and so from Slutsky's Theorem

$$\sqrt{n}\frac{\bar{Y}-\mu}{s} = \frac{\sqrt{n}(\bar{Y}-\mu)/\sigma}{s/\sigma} \xrightarrow{\text{d}} \frac{\mathcal{N}(0,1)}{1} = \mathcal{N}(0,1)$$

as $n \to \infty$.

Exercises

2.1. This is the time-honored π-estimation by Monte Carlo exercise, with a twist to estimate the accuracy of the estimate. Let $\boldsymbol{X} = (X_1, X_2)$ be uniformly distributed in the unit square $[0,1] \times [0,1]$. Let $Y = f(\boldsymbol{X})$ where

$$f(\boldsymbol{x}) = \begin{cases} 1, & x_1^2 + x_2^2 \leqslant 1 \\ 0, & \text{else.} \end{cases}$$

a) Use a Monte Carlo sample of size $n = 1000$ to produce an approximate 95% confidence interval for $\mu = \mathbb{E}(Y)$ based on the CLT. Translate your confidence interval into a confidence interval for π.

b) Repeat the previous simulation 1000 times independently, and report how many of the confidence intervals you get actually contained π.

2.2. One way to define a poverty line is to say it is half of the median income. The low income proportion $\text{LIP}(\alpha, \beta)$ is the fraction of incomes that are below α times the β quantile of incomes. Taking $\alpha = \beta = 1/2$ yields the fraction of incomes corresponding to poverty. For example, if the income distribution is $\mathbf{U}[0, 10]$, then the median income is 5, half of that is 2.5 and 25% of incomes are below 2.5 for an LIP of 0.25. More realistic income distributions have a long tail to the right. Suppose that income Y has a log-normal distribution: $Y = \exp(\mu + \sigma Z)$ where $Z \sim \mathcal{N}(0,1)$ and $\sigma > 0$.

a) Explain why the value of μ does not affect the LIP for the log normal distribution.

b) For $\sigma = 0.7$, find $\text{LIP}(1/2, 1/2)$, the fraction of incomes below half the median. This can be done using quantiles of the $\mathcal{N}(0,1)$ distribution (i.e., Monte Carlo is not required).

c) Suppose that $\sigma = 0.7$ and we get a sample of $n = 101$ incomes $Y_1, \ldots, Y_{101}$, find their median, halve it and record the fraction of Y_i below half the median. What then will be the mean and variance of our estimate? What is the chance that our estimate is below $0.75 \times \text{LIP}(1/2, 1/2)$? What is the chance that is above $1.25 \times \text{LIP}(1/2, 1/2)$? Use Monte Carlo to answer these questions. It is enough to repeat the sampling 10,000 times. Show a histogram of your LIP values.

For this exercise you need to sample independent $\mathcal{N}(0,1)$ random variables. If your computing environment does not supply them, then look into §4.3 or §4.6 for algorithms to convert $\mathbf{U}(0,1)$ random variables to $\mathcal{N}(0,1)$.

2.3. Prove that update equations in (2.18) correctly yield S_n and $\hat{\mu}_n$.

2.4. In this problem, we learn how hard it is to break the variance formula (2.17). The data are simply $y_i = \Delta + i/n$ where $\Delta \geqslant 0$ and $n > 0$. The true value of $\hat{\sigma}^2$ depends on n but not on Δ and is slightly less than $1/12$. The numerics are easy when $\Delta = 0$ and hard for large Δ.

a) For $n = 10$, 100, and 1000 and using $\Delta = 0$, report twelve times the value of (2.17).

b) For $n = 1000$ and $\Delta = 2^j$ for integers j from 0 to $J = 100$ compute $12\hat{\sigma}^2$ using (2.17). Report the smallest j (if any) for which each of the following happen: $|12\hat{\sigma}^2 - 1| > 1$, $\hat{\sigma}^2 < 0$, $\hat{\sigma}^2 = 0$.

c) Repeat the previous part, using $12S_n/n$ via (2.18). Say whether (2.18) appears to be better than, worse than, or about as good as (2.17). Does (2.18) appear unbreakable?

d) For the benefit of your grader, identify the programming language or package, computer architecture, and operating system on which you got your answers.

2.5. Let Y_i be IID random variables with mean μ, variance $\sigma^2 > 0$ and finite kurtosis $\kappa = \mathbb{E}((Y-\mu)^4)/\sigma^4 - 3$. It is well known that $\mathrm{Var}(s^2) = \sigma^4(2/(n-1) + \kappa/n)$. For even n, an alternative estimate $\widetilde{\sigma}^2 = (1/n)\sum_{i=1}^{n/2}(Y_{2i-1} - Y_{2i})^2$ was discussed on page 22.

a) Show that $\mathbb{E}(\widetilde{\sigma}^2) = \sigma^2$.

b) Show that $\mathrm{Var}(\widetilde{\sigma}^2) = \sigma^4(4+\kappa)/n$.

Note: if $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ then $\kappa = 0$ and $\mathrm{Var}(\widetilde{\sigma}^2) \doteq 2\mathrm{Var}(s^2)$, for large n. If Y is heavy-tailed so that $\kappa \gg 4$, then $\mathrm{Var}(\widetilde{\sigma}^2) \doteq \mathrm{Var}(s^2)$. So except for short-tailed Y with $\kappa < 0$, the estimate $\widetilde{\sigma}^2$ has accuracy comparable to s^2 with between $n/2$ and n observations.

2.6. Here we look at the effect of using $\widetilde{\sigma}^2$ of (2.19) versus s^2 in a confidence interval for μ. To make things simple we consider $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. As the previous exercise shows, $\widetilde{\sigma}^2$ has about twice the variance as s^2, for large n, in this case.

For normally distributed data, the standard estimate leads to a 99% confidence interval

$$\bar{Y} \pm T^{0.995}_{(n-1)}\, s/\sqrt{n}$$

where $T^{0.995}_{(k)}$ is 99.5th percentile of Student's t distribution on k degrees of freedom and $(n-1)s^2/\sigma^2 \sim \chi^2_{(n-1)}$. Using the modified variance estimate $\widetilde{\sigma}^2$, we would instead compute a 99.5% confidence interval

$$\bar{Y} \pm T^{0.995}_{(n/2)}\, \widetilde{\sigma}/\sqrt{n}.$$

where $(n/2)\widetilde{\sigma}^2/\sigma^2 \sim \chi^2_{(n/2)}$.

Let W be the width of the standard interval and $\widetilde{W}$ be the width of the modified one.

a) Find a formula for $\mathbb{E}(\widetilde{W}^2)/\mathbb{E}(W^2)$ as a function of n.

b) What values do you get for $n = 10^r$ and $r \in \{1, 2, 3, 4\}$?

2.7. Suppose that a truly enormous Monte Carlo simulation is being done. It has $n = 10^{15}$ and the machine precision is 10^{-16}. Then the relative error in s^2 computed by updating is known to be below $10^{-1}\sqrt{1+\hat{\mu}^2/\hat{\sigma}^2}$. This bound only guarantees between 0 and 1 digit of accuracy in the computed value of s^2.

Somebody proposes using $\hat{\mu}_n$ based on all n simulation values and computing $\hat{\sigma}^2_{\widetilde{n}}$ based on only the first $\widetilde{n} = 10^7$ of the simulation values. Is this a workable idea? If so, explain how to compute an approximate 99% confidence interval for μ using $\hat{\mu}_n$ and $\hat{\sigma}^2_{\widetilde{n}}$. If not, describe what goes wrong statistically, numerically, or both.

2.8. Suppose that we sample $Y_i \in \{0, 1\}$ independently for $i = 1, \dots, n$ using $n = 10{,}000$ and get $Y_i = 1$ exactly 17 times. Using equations (2.21) and (2.22) create an exact 99% confidence interval for $p = \mathbb{E}(Y)$.

2.9. Show that equation (2.28) simplifies to $(1/n)\mathbb{E}((Y-\theta X)^2)/\mu_x^2$ for the ratio estimator $\theta = f(\mathbb{E}(X), \mathbb{E}(Y)) = \mathbb{E}(Y)/\mathbb{E}(X)$.

2.10. Simplify equation (2.30) for the special case of a ratio $\theta = \mathbb{E}(YA)/\mathbb{E}(A)$. Here $Y = f(X)$ and $A(X) \in \{0, 1\}$. That is YA plays the role of Y, and A plays that of X in $\theta = \mathbb{E}(Y)/\mathbb{E}(X)$.

2.11. Let $Y_1, \dots, Y_T$ be independent and identically distributed. Let $S_h = \bar{Y}/s$, where

$$\bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t, \quad \text{and} \quad s^2 = \frac{1}{T-1}\sum_{t=1}^{T}(Y_t - \bar{Y})^2.$$

Derive the form of the delta method variance estimate (2.40) for the statistic S_h. You may replace $T - 1$ by T in the definition of s^2. It may simplify your derivation to use the variance formula (2.17).

If Y_t is the return to a stock in time period t, minus the return of a benchmark item, then S_h is the ***Sharpe ratio*** for that stock.

2.12. Suppose that we do a Monte Carlo simulation with $n = 10{,}000$ and we want to estimate $Q^{0.05}$, the 5'th percentile of Y. Find integers L and R so that $[Y_{(L)}, Y_{(R)}]$ is a conservative 99% confidence interval for $Q^{0.05}$.

2.13. Prove that Algorithm 2.1 correctly implements equation (2.26), for the quantile confidence interval problem. This exercise depends on knowing how inverse CDFs are defined. You may read ahead into §4.1 for the definition.

2.14. In this exercise, we look at what happens when Monte Carlo is applied for a mean that does not exist.

When we arrive at the transit station, we could take either a bus or a taxi. The waiting time for a taxi is $X \sim \text{Exp}(1)$, while the waiting time for a bus is $Y \sim \mathbf{U}(0, 1)$ independently of X. These are the familiar exponential and uniform distributions respectively. If necessary, see Chapter 4 for descriptions.

a) Prove that $\mathbb{E}(Y/X) = \infty$, so that by this measure the bus is infinitely worse than the taxi service.

b) Counter the previous part by proving that $\mathbb{E}(X/Y) = \infty$ too.

c) Simulate $N = 10{,}000$ independent (X_i, Y_i) pairs and, for $1 \leqslant n \leqslant N$, let $\hat{\mu}_n = (1/n)\sum_{i=1}^{n} Y_i/X_i$. Do this simulation 10 independent times and plot all 10 curves of $\hat{\mu}_n$ versus n. (Use your judgment as to whether a logarithmic scale is better for the vertical axis, and/or whether to truncate the vertical axis at some large value and whether to overlay the curves or use separate plot panels for each.) What sign, if any, do these plots give that $\mathbb{E}(Y/X)$ might not be finite?

d) Plot the upper and lower limits of the CLT-based 99% confidence interval formula (2.13) versus sample sizes $2 \leqslant n \leqslant N$, for one of the 10 curves you generated.

As noted in the text these intervals have essentially zero coverage when $\mu = \infty$. When the CLT holds the confidence interval widths scale proportionally to $1/\sqrt{n}$. How do the widths behave for your example?

This exercise assumes that you have $\mathbf{U}(0,1)$ random variables available. To sample $Y \sim \mathrm{Exp}(1)$ you can take $Y = -\log(U)$ where $U \sim \mathbf{U}(0,1)$.

2.15. This exercise uses Monte Carlo to see patterns in the life times of long-lived comets as described in Example 2.5. Simulate 10,000 comets starting with $x_0 = -1$ and continuing until they either get positive energy (leaving the solar system) or complete 10^5 orbits, whichever happens first.

a) Give the Agresti confidence interval, based on your data, for the probability that a comet will make more than 10^5 orbits. Do the same for 10^k orbits, $k \in \{1, 2, 3, 4\}$.

b) Give an estimate and an Agresti 99% confidence interval for the probability that a comet survives more than 10^{k+1} orbits given that it has survived more than 10^k orbits where $1 \leqslant k \leqslant 4$.

c) For those comets observed to leave the solar system, show the histogram of the number of orbits observed and make a plot to show how the observed life time is related to the number of orbits.

d) Find the 10 comets with the longest observed lifetimes, whether or not they were observed to leave the solar system. Plot their energy trajectories: $|x_j|$ versus j where $j = 0, 1, \ldots$ is the number of orbits made. Do you find that $|x_j|$ increases linearly or nonlinearly but monotonically, or haphazardly for your 10 comets?

e) Estimate the mean lifetime for comets that stay for 1000 orbits or fewer, and give a 99% confidence interval.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, New York, 2nd edition.

Boyle, P. P. (1977). Options: A Monte Carlo approach. *Journal of Financial Economics*, 4(3):323–338.

Chan, T. F., Golub, G. H., and LeVeque, R. J. (1983). Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, 37(3):242–247.

Chowdhury, D., Santen, L., and Schadschneiter, A. (2000). Statistical physics of vehicular traffic and some related problems. *Physics reports*, 329:199–329.

Chung, K.-L. (1974). *A course in probability theory*. Academic press, New York.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer, New York.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B*, 16(2):175–185.

Gelfand, A. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Ghosh, B. (1951). Random distances within a rectangle and between two rectangles. *Bulletin of the Calcutta Mathematical Society*, 43(1):17–24.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Chemical Physics*, 81(25):2340–2361.

Gumbel, E. J. (1939). La probabilité des hypothèses. *Comptes Rendus de l'Académie des Sciences (Paris)*, 209:645–647.

Hahn, G. J. and Meeker, W. Q. (1991). *Statistical intervals: a guide for practitioners*. Wiley, New York.

Hall, P. G. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hammersley, J. M. (1961). On the statistical loss of long period comets from the solar system, ii. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pages 17–78.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo methods*. Methuen, London.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365.

Kajiya, J. T. (1988). The rendering equation. In *SIGGRAPH '86 Conference Proceedings*, volume 20, pages 143–150. Association for Computing Machinery.

Kalos, M. H. and Whitlock, P. A. (2008). *Monte Carlo Methods*. Wiley, New York.

Kelvin, L. (1901). Nineteenth century clouds over the dynamical theory of heat and light. *Philosophical Magazine*, 6(2):1–40.

Kirkpatrick, S., Gelatt Jr., C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Knight, K. (2000). *Mathematical Statistics*. Chapman & Hall/CRC, Boca Raton, FL.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, 3rd edition.

Marsaglia, G., Narasimhan, B., and Zaman, A. (1990). The distance between random points in rectangles. *Communications in Statistics - Theory and Methods*, 19(11):4199–4212.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1091.

Nagel, K. and Schreckenberg, M. (1992). A cellular automaton model for freeway traffic. *Journal de Physique I*, 2(12):2221–2229.

Peng, L. (2004). Empirical-likelihood-based confidence interval for the mean with a heavy-tailed distribution. *Annals of Statistics*, 32(3):1192–1214.

Qin, H. and Li, L. (2001). Comparison of variance estimators for the ratio estimator based on small sample. *Acta Mathematicae Applicatae Sinica*, 17(4):449–456.

Richtmyer, R. D. (1952). The evaluation of definite integrals, and a quasi-Monte Carlo method based on the properties of algebraic numbers. Technical Report LA-1342, University of California.

Robbins, D. (1978). Average distance between two points in a box. *American Mathematical Monthly*, 85(4):278.

Student (1908). The probable error of a mean. *Biometrika*, 6(1):1–25.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Tippett, L. H. C. (1927). *Random sampling numbers*. Cambridge University Press, Cambridge.

Tocher, K. D. and Owen, D. G. (1960). The automatic programming of simulations. In Banbury, J. and Maitland, J., editors, *2nd International Conference on Operational Research*, pages 60–68, London. English Universities Press.

Weisstein, E. W. (2005). Cube line picking, Mathworld–a Wolfram web resource. `http://mathworld.wolfram.com/CubeLinePicking.html`.

Welford, B. P. (1962). Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420.

Weyl, H. (1914). Über ein problem aus dem gebiete der diophantischen approximationen. *Nachrichten der Akademie der Wissenschaften in Göttingen. II. Mathematisch-Physikalische Klasse*, pages 234–244.

Weyl, H. (1916). Über die gleichverteilung von zahlen mod. eins. *Mathematische Annalen*, 77:313–352.