A Friendly Introduction to Probability

A Simulation-based Approach with Python

Kevin Ross

Sep 26, 2025

Table of contents

Preface

Why study probability *and simulation*?

Why study probability?

- Probability is the study of uncertainty, and life is uncertain
- Probability is used in a wide variety of fields, including: statistics, physics, engineering, biology, medicine, finance, actuarial science, political science, law, sports , …
- Many topics and problems in probability are frequently misunderstood and sometimes counter intuitive, so it's worthwhile to take a careful study
- "Probabilistic thinking" is an important component of statistical literacy (e.g. how to assess risk when making decisions)
- Probability provides the foundation for many important statistical concepts and methods such as p-values and confidence intervals

Why use **simulation** to study probability?

- Many concepts encountered in probability can seem esoteric; simulation helps make them more concrete.
- Simulation provides an effective tool for analyzing probability models and for exploring effects of changing assumptions
- Simulation can be used to check analytical solutions
- Simulation is often the best or only method for investigating many problems which are too complex to solve analytically
- Simulation allows for *statistical* approaches to solving probability problems (e.g., treat the simulated values as data to summarize and analyze)
- Simulation-based reasoning is an important component of statistical literacy (e.g., understanding a p-value via simulation)
- Many statistical procedures employ simulation-based methods (e.g. bootstrapping)

Learn by doing

There are many examples in this book. Examples are used to both motivate new topics and to help you practice your understanding of the material. You should attempt the examples on your own before reading the solutions. To encourage you to do so, the solutions have been hidden.

Solution (click to expand)

You can reveal the solution by clicking on the "Solution (click to expand)" button.

Don't do what Donny Don't does

Some of the examples in this book feature a character named Donny Don't. The moral of these examples is (usually) "Don't do what Donny Don't does". (This is a Simpson's reference.) Donny represents a student who makes many of the mistakes commonly made by students studying probability. The idea of these problems is for you to learn from the common mistakes that Donny makes, by identifying why he is wrong and by helping him understand and correct his mistakes. But be careful: sometimes Donny is right!

Learning objectives

At the completion of this book, you should be able to:

- Interpret conditional and unconditional probabilities and expected values
- Identify coherent probability models, events and random variables
- Apply common probability models in real contexts
- Design simulation studies to investigate random phenomena
- Analyze simulation results
- Solve probability problems using mathematical properties and tools
- Describe distributions of random variables
- Construct representations of distributions

Symbulate

This book uses the Python package Symbulate which provides a user friendly framework for conducting simulations involving probability models. The syntax of Symbulate reflects the "language of probability" and makes it intuitive to specify, run, analyze, and visualize the results of a simulation. In Symbulate, probability spaces, events, random variables, and random processes are symbolic objects which can be manipulated, independently of their simulated realizations. Symbulate's consistency with the mathematics of probability reinforces understanding of probabilistic concepts. The article (Ross and Sun 2019) discusses Symbulate and its features in more detail.

The best way to interact with Symbulate is through Google Colab or Jupyter notebooks. A notebook is organized by cells which contain text or code that can be run interactively with output displayed after the cell.

Symbulate can be run online in a Colab notebook by including the following line in the first cell.

```
pip install git+https://github.com/kevindavisross/symbulate
```

Symbulate can also be used in RMarkdown or Quarto documents, if you install the package on your device. To install Symbulate on your own computer, it is recommended that you first install the Anaconda distribution, which is a Python environment with many scientific packages installed (including all of the packages that Symbulate is built on). After installing Anaconda, you can install Symbulate using the `pip` command above.

You should always include the following command once in each notebook to import Symbulate during a Python session.

```
from symbulate import *
```

The Symbulate command `plot()` produces graphics. These graphics can be customized (by changing axis limits, adding titles, legends, etc) using Matplotlib, and in particular the `pyplot` method, which can be imported by including the lines

```
import matplotlib
import matplotlib.pyplot as plt
```

Colab or Jupyter notebooks provide a natural interface for Symbulate. The code in this book matches as closely as possible the commands that would be entered into cells in a notebook. However, certain commands that appear throughout the book are needed only to properly produce the output in this book, and not if working directly in notebooks (e.g., some `print` statements, instances of `plt.figure()` or `plt.show()`)

About this book

This book was created with Quarto.

1 What is Probability?

We hear the word "probability" often. Here are just a few quotes from recent online articles which mention probability.

- Researchers say the *probability* of living past 110 is on the rise (CNBC, July 17, 2021)
- Forecasters now have a model that can predict the *probability* of rip currents up to six days out. (CNN, July 18, 2021)
- A Week of negatives increases *probability* of stock market correction. (Forbes, July 10, 2021)
- Less than 1% *probability* that Earth's energy imbalance increase occurred naturally (Science Daily, July 28, 2021)
- Basketball Reference's Hall of Fame *probability* model already has Giannis at 67.9 percent. (Bleacher Report, July 23, 2021)
- Study suggests that the rate of global warming increases the *probability* of extreme temperatures. (NPR, July 29, 2021)
- We anticipate an above-average *probability* for major hurricanes making landfall along the continental United States coastline and in the Caribbean. (Weather Channel, June 1, 2021)
- Scientists fine-tune *odds* of asteroid Bennu hitting Earth through 2300 with NASA probe's help (Space.com, Sept 3, 2021)

You have some familiarity with the words "probability", "chance", "odds", or "likelihood" from everyday life. But what do we really mean when talk about "probability"?

This chapter provides a brief but non-technical introduction to randomness and probability. Many of the topics introduced in this chapter will be covered in much more detail in later chapters.

1.1 Randomness

A wide variety of situations involve probability. Consider just a few examples.

- The probability that you roll doubles in a turn of a board game.
- The probability you win the next Powerball lottery if you purchase a single ticket, 4-8-15-16-42, plus the Powerball number, 23.
- The probability that a randomly selected Cal Poly student is a California resident.

- The probability that the high temperature in San Luis Obispo, CA tomorrow is above 90 degrees F.
- The probability that Hurricane Martin makes landfall in the U.S in 2028.
- The probability that the Philadelphia Eagles win the next Superbowl.
- The probability that the Republican candidate wins the 2032 U.S. Presidential Election.
- The probability that extraterrestrial life currently exists somewhere in the universe.
- The probability that Alexander Hamilton actually wrote 51 of The Federalist Papers. (The papers were published under a common pseudonym and authorship of some of the papers is disputed.)
- The probability that you ate an apple on April 17, 2019.

Example 1.1. What is one feature that all of the situations have in common? How are the situations above similar, and how are they different? Is the interpretation of "probability" the same in all situations? The goal here is to just think about these situations, and not to compute any probabilities (or to even think about how you would).

Solution (click to expand)

Solution 1.1. This example is intended to motivate discussion, so you might have thought of some other ideas we don't address here. That's good! And some of the things you considered might come up later in the book. Here are a few observations.

The one feature that all of the situations have in common is *uncertainty*. Sometimes the uncertainty arises from a physical phenomenon that can result in multiple potential outcomes, like rolling dice or drawing the winning Powerball number. In other cases, there is uncertainty because there will be only one outcome but it is in the future, like tomorrow's high temperature or the result of the next Superbowl. But there can also be uncertainty about the past: there are some Federalist papers for which the author is unknown, and you probably don't know for sure whether or not you ate an apple on April 17, 2019.

Whenever there is uncertainty, it is reasonable to consider the relative likelihood or plausibility of possibilities. For example, even though you don't know for certain whether you ate an apple on April 17, 2019, you might think the probability is high if you're usually an apple-a-day person (or you were in 2019). We don't know for sure what team will win the next Superbowl, but we might think that the Eagles are more likely than the Cleveland Brows to be the winner.

While all of the situations in this example involve uncertainty, it seems that there are different "types" of uncertainty. Even though we don't know which side a die will land on, the notion of "fairness" implies that the sides are "equally likely". Likewise, there are some rules to how the Powerball drawing works, and it seems like these rules should determine the probability of drawing that particular winning number.

However, there aren't any specific "rules of uncertainty" that govern whether or not you ate an apple on April 17, 2019. You either did or you didn't, but that doesn't mean these two possibilities are necessarily equally likely or plausible. Regarding the Superbowl, of course there are rules that govern the NFL season and playoffs, but there are no "rules of uncertainty" that tell us precisely how likely any particular team is to win any particular game, let alone how likely a team is to advance to and win the Superbowl.

It also seems that there are different interpretations of probability. Given that a six-sided die is fair, we might all agree that the probability that it lands on any particular side is 1/6. Similarly, given the rules of the Powerball lottery, we might all agree on the probability that a drawing results in a particular winning number. However, there isn't necessarily consensus about what the high temperature will be in San Luis Obispo tomorrow; different weather prediction models, forecasters, or websites might provide different values for the probability that the high temperature will be above 90 degrees Fahrenheit. Similarly, Superbowl odds might vary by source. Situations like tomorrow's weather or the Superbowl where there is no consensus about the "rules of uncertainty" require some subjectivity in determining probabilities.

Finally, some of these situations are naturally repeatedable. We could (in principle) roll a pair of dice many times and see how often we get doubles, or repeat the Powerball drawing over and over to see how the winning numbers behave. However, many of these situations involve something that only happens once, like the next Superbowl, tomorrow, or April 17, 2019. Even when the phenomenon happens only once in reality, we can still develop models of what *might* happen if we were to *hypothetically* repeat the phenomenon many times. For example, meteorologists use historical data and meteorological models to forecast many potential paths of a hurricane.

The subject of probability concerns *random* phenomena.

Definition 1.1. A phenomenon is **random** if there are multiple potential possibilities, and there is *uncertainty* about which possibility is realized. Uncertainty is understood in broad terms, and in particular does not only concern future occurrences.

Some phenomena involve physical randomness, like flipping coins, rolling dice, drawing Powerballs at random from a bin, or random digit dialing. In many other situations randomness just vaguely reflects uncertainty. We will refer to as "random" any scenario that involves a reasonable degree of uncertainty.

In this book, "random" and "uncertain" are synonyms. Unfortunately, some of the everyday meanings of "random", like "haphazard" or "unexpected", are contrary to what we mean by "random" in this book. For example, we would consider Steph Curry attempting a free throw to be a random phenomenon because we're not certain if he'll make it or miss it; but we would not consider this process to be haphazard or unexpected.

Random does *not* necessarily mean equally likely. In a random phenomenon, certain outcomes or events might be more or less likely than others. For example,

- About 84% of students at Cal Poly are California residents, so it's more likely than not that a randomly selected Cal Poly student is a California resident.
- Not all NFL teams are equally likely to win the next Superbowl.

Uncertainty is not something to be feared, and randomness is often desirable. In particular, many statistical applications often employ the planned use of randomness with the goal of collecting "good" data. For example,

- *Random selection* involves selecting a sample of individuals at random from a population (e.g., via random digit dialing), with the goal of selecting a representative sample.

- *Random assignment* involves assigning individuals at random to groups (e.g., in a randomized experiment), with the goal of constructing groups that are similar in all aspects so that the effect of a treatment (like a new vaccine) can be isolated.

1.1.1 Exercises

Exercise 1.1. For each of the following, provide examples of random phenomenon that fit the description. Try to think of examples that are interesting to you personally!

1. Just two possible outcomes, but they are not equally likely.
2. Physically repeatable (at least in principle).
3. Well defined "rules of randomness".
4. Involves subjectivity in determining probabilities.
5. Involves uncertainty about the future.
6. Involves uncertainty about the present or past.
7. Associated with the planned use of randomness in a particular statistical study.

1.2 Interpretations of probability

The **probability** of an event associated with a random phenomenon is a number in the interval $[0, 1]$ measuring the event's likelihood, degree of uncertainty, or relative plausibility. A probability can take any value in the continuous scale from 0 to 1, and can be reported either as a decimal (e.g., 0.305) or as a percent (e.g., 30.5%).

A few examples of probabilities:

- The probability that a fair coin lands on heads 5 times in 10 flips is 0.246.
- A group of people all put their names in a hat for a Secret Santa gift exchange; the probability that at least one person in the group draws their own name is 0.632.
- The probability that a randomly selected full term baby weighs more than 4000 grams at birth is 0.09.

- The probability that a magnitude 5+ earthquake occurs somewhere in the world within the next 48 hours is 0.96.
- According to FiveThirtyEight as of Nov 8, 2016, the probability that Donald Trump would win the 2016 U.S. Presidential Election was 0.286.

Throughout this book we will see many methods for computing and approximating probabilities such as these. But given the value of a probability, what does it mean? For example, what does it mean for there to be a "30% chance of rain tomorrow"? Just as there are various types of randomness, there are a few ways of interpreting probability, most notably, *long run relative frequency* and *subjective probability*.

1.2.1 Long run relative frequency

One of the oldest documented[1] problems in probability is the following: If three fair six-sided dice are rolled, what is more likely—a sum of 9 or a sum of 10? Let's try to answer this question by simply rolling dice and seeing what happens. Roll three fair six-sided dice, find the sum, and repeat; then see how often we get a sum of 9 versus a sum of 10. Table **??** displays the results of a few repetitions. We encourage you to try this out on your own now; of course, your results will naturally be different from ours.

Table 1.1: Results of 10 sets of three rolls of a fair six-sided die.

Repetition	First roll	Second roll	Third roll	Sum
1	3	6	3	12
2	1	2	4	7
3	4	2	4	10
4	2	2	1	5
5	4	6	1	11
6	5	1	2	8
7	3	1	3	7
8	5	5	6	16
9	5	6	3	14
10	4	5	2	11

A sum of 9 occurred in 0 repetitions and a sum of 10 occurred in 1 repetition. We see that a sum of 10 occurred more frequently than a sum of 9, but our results should not be very convincing. After all, we only performed 10 repetitions and your results are probably different than ours. We can get a much better picture by performing many, many repetitions. This

[1]The Grand Duke of Tuscany posed this problem to Galileo, who published his solution in 1620. However, unbeknownst to Galileo, the same problem had been solved almost 100 years earlier by Gerolamo Cardano, one of the first mathematicians to study probability (David 1955).