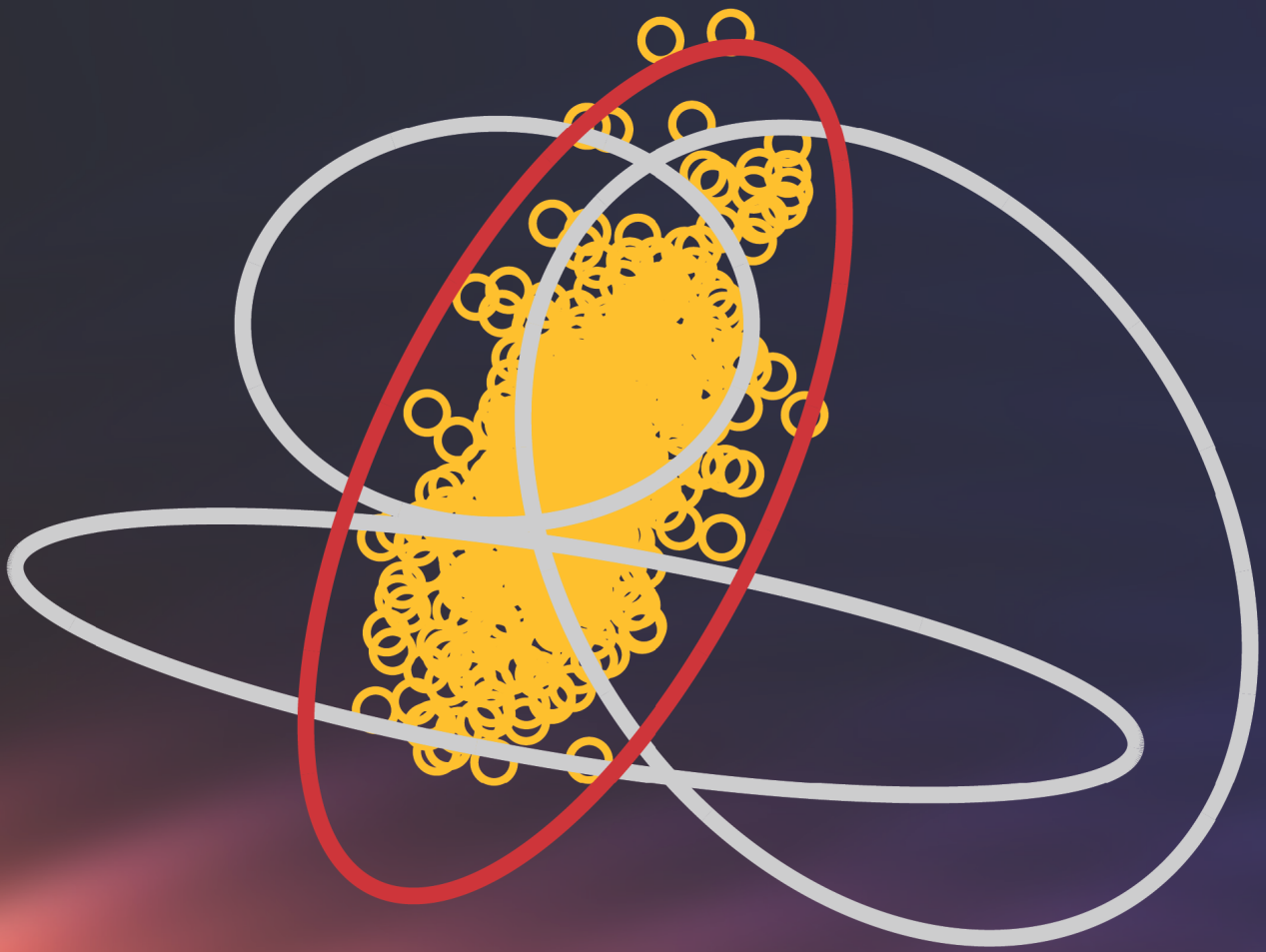
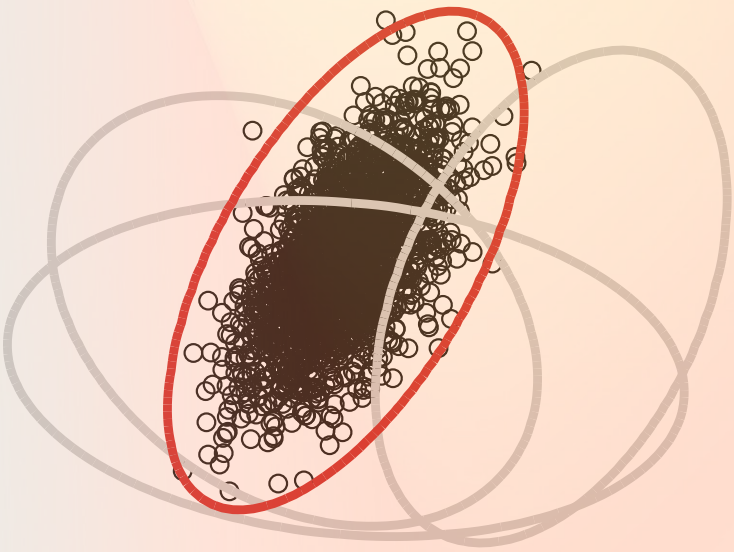
Introduction to PROBABILITY for DATA SCIENCE



Stanley H. Chan

Introduction to Probability
for
Data Science

Stanley H. Chan
Purdue University



Published in the United States of America by
Michigan Publishing
Manufactured in the United States of America

ISBN 978-1-60785-746-4 (hardcover)
ISBN 978-1-60785-747-1 (electronic)

To Vivian, Joanna, and Cynthia Chan

And ye shall know the truth, and the truth shall make you free.

John 8:32

Preface

This book is an introductory textbook in undergraduate probability. It has a mission: to spell out the *motivation*, *intuition*, and *implication* of the probabilistic tools we use in science and engineering. From over half a decade of teaching the course, I have distilled what I believe to be the core of probabilistic methods. I put the book in the context of data science to emphasize the inseparability between data (computing) and probability (theory) in our time.

Probability is one of the most interesting subjects in electrical engineering and computer science. It bridges our favorite engineering principles to the practical reality, a world that is full of uncertainty. However, because probability is such a mature subject, the undergraduate textbooks alone might fill several rows of shelves in a library. When the literature is so rich, the challenge becomes how one can pierce through to the insight while diving into the details. For example, many of you have used a normal random variable before, but have you ever wondered where the "bell shape" comes from? Every probability class will teach you about flipping a coin, but how can "flipping a coin" ever be useful in machine learning today? Data scientists use the Poisson random variables to model the internet traffic, but where does the gorgeous Poisson equation come from? This book is designed to fill these gaps with knowledge that is essential to all data science students.

This leads to the three goals of the book. (i) Motivation: In the ocean of mathematical definitions, theorems, and equations, why should we spend our time on this particular topic but not another? (ii) Intuition: When going through the derivations, is there a geometric interpretation or physics beyond those equations? (iii) Implication: After we have learned a topic, what new problems can we solve?

The book's intended audience is undergraduate juniors/seniors and first-year graduate students majoring in electrical engineering and computer science. The prerequisites are standard undergraduate linear algebra and calculus, except for the section about characteristic functions, where Fourier transforms are needed. An undergraduate course in signals and systems would suffice, even taken concurrently while studying this book.

The length of the book is suitable for a two-semester course. Instructors are encouraged to use the set of chapters that best fits their classes. For example, a basic probability course can use Chapters 1-5 as its backbone. Chapter 6 on sample statistics is suitable for students who wish to gain theoretical insights into probabilistic convergence. Chapter 7 on regression and Chapter 8 on estimation best suit students who want to pursue machine learning and signal processing. Chapter 9 discusses confidence intervals and hypothesis testing, which are critical to modern data analysis. Chapter 10 introduces random processes. My approach for random processes is more tailored to information processing and communication systems, which are usually more relevant to electrical engineering students.

Additional teaching resources can be found on the book's website, where you can

find lecture videos and homework videos. Throughout the book you will see many "practice exercises", which are easy problems with worked-out solutions. They can be skipped without loss to the flow of the book.

Acknowledgements: If I could thank only one person, it must be Professor Fawwaz Ulaby of the University of Michigan. Professor Ulaby has been the source of support in all aspects, from the book's layout to technical content, proofreading, and marketing. The book would not have been published without the help of Professor Ulaby. I am deeply moved by Professor Ulaby's vision that education should be made accessible to all students. With textbook prices rocketing up, the EECS free textbook initiative launched by Professor Ulaby is the most direct response to the publishers, teachers, parents, and students. Thank you, Fawwaz, for your unbounded support — technically, mentally, and financially. Thank you also for recommending Richard Carnes. The meticulous details Richard offered have significantly improved the fluency of the book. Thank you, Richard.

I thank my colleagues at Purdue who had shared many thoughts with me when I taught the course (in alphabetical order): Professors Mark Bell, Mary Comer, Saul Gelfand, Amy Reibman, and Chih-Chun Wang. My teaching assistant I-Fan Lin was instrumental in the early development of this book. To the graduate students of my lab (Yiheng Chi, Nick Chimitt, Kent Gauen, Abhiram Gnanasambandam, Guanzhe Hong, Chengxi Li, Zhiyuan Mao, Xiangyu Qu, and Yash Sanghvi): Thank you! It would have been impossible to finish the book without your participation. A few students I taught volunteered to help edit the book: Benjamin Gottfried, Harrison Hsueh, Dawoon Jung, Antonio Kincaid, Deepak Ravikumar, Krister Ulvog, Peace Umoru, Zhijing Yao. I would like to thank my Ph.D. advisor Professor Truong Nguyen for encouraging me to write the book.

Finally, I would like to thank my wife Vivian and my daughters, Joanna and Cynthia, for their love, patience, and support.

Stanley H. Chan, *West Lafayette, Indiana*

May, 2021

Companion website:

`https://probability4datascience.com/`

Contents

Chapter 1

Mathematical Background

"Data science" has different meanings to different people. If you ask a biologist, data science could mean analyzing DNA sequences. If you ask a banker, data science could mean predicting the stock market. If you ask a software engineer, data science could mean programs and data structures; if you ask a machine learning scientist, data science could mean models and algorithms. However, one thing that is common in all these disciplines is the concept of **uncertainty**. We choose to learn from data because we believe that the latent information is embedded in the data — unprocessed, contains noise, and could have missing entries. If there is no randomness, all data scientists can close their business because there is simply no problem to solve. However, the moment we see randomness, our business comes back. Therefore, data science is the subject of making decisions in uncertainty.

The mathematics of analyzing uncertainty is **probability**. It is *the* tool to help us model, analyze, and predict random events. Probability can be studied in as many ways as you can think of. You can take a rigorous course in probability theory, or a "probability for dummies" on the internet, or a typical undergraduate probability course offered by your school. This book is different from all these. Our goal is to tell you *how things work* in the context of data science. For example, why do we need those three axioms of probabilities and not others? Where does the "bell shape" Gaussian random variable come from? How many samples do we need to construct a reliable histogram? These questions are at the core of data science, and they deserve close attention rather than sweeping them under the rug.

To help you get used to the pace and style of this book, in this chapter, we review some of the very familiar topics in undergraduate algebra and calculus. These topics are meant to warm up your mathematics background so that you can follow the subsequent chapters. Specifically, in this chapter, we cover several topics. First, in Section 1.1 we discuss infinite series, something that will be used frequently when we evaluate the expectation and variance of random variables in Chapter 3. In Section 1.2 we review the Taylor approximation, which will be helpful when we discuss continuous random variables. Section 1.3 discusses integration and reviews several tricks we can use to make integration easy. Section 1.4 deals with linear algebra, aka matrices and vectors, which are fundamental to modern data analysis. Finally, Section 1.5 discusses permutation and combination, two basic techniques to count events.

1.1 Infinite Series

Imagine that you have a **fair coin**. If you get a tail, you flip it again. You do this repeatedly until you finally get a head. What is the probability that you need to flip the coin three times to get one head?

This is a warm-up exercise. Since the coin is fair, the probability of obtaining a head is $\frac{1}{2}$. The probability of getting a tail followed by a head is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Similarly, the probability of getting two tails and then a head is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$. If you follow this logic, you can write down the probabilities for all other cases. For your convenience, we have drawn the first few in **Figure 1.1**. As you have probably noticed, the probabilities follow the pattern $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots\}$.

Figure 1.1: Suppose you flip a coin until you see a head. This requires you to have $N-1$ tails followed by a head. The probability of this sequence of events are $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, ..., which forms an infinite sequence.

We can also summarize these probabilities using a familiar plot called the **histogram** as shown in **Figure 1.2**. The histogram for this problem has a special pattern, that every value is one order higher than the preceding one, and the sequence is infinitely long.

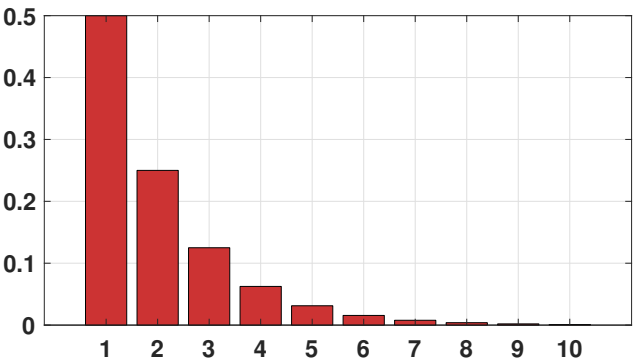


Figure 1.2: The histogram of flipping a coin until we see a head. The x-axis is the number of coin flips, and the y-axis is the probability.

Let us ask something harder: On average, if you want to be 90% sure that you will get a head, what is the minimum number of attempts you need to try? Five attempts? Ten attempts? Indeed, if you try ten attempts, you will very likely accomplish your goal. However, this would seem to be overkill. If you try five attempts, then it becomes unclear whether you will be 90% sure.

This problem can be answered by analyzing the sequence of probabilities. If we make two attempts, then the probability of getting a head is the sum of the probabilities for one attempt and that of two attempts:

$$\mathbb{P}[\text{success after 1 attempt}] = \frac{1}{2} = 0.5$$
$$\mathbb{P}[\text{success after 2 attempts}] = \frac{1}{2} + \frac{1}{4} = 0.75$$

Therefore, if you make 3 attempts or 4 attempts, you get the following probabilities:

$$\mathbb{P}[\text{success after 3 attempts}] = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = 0.875$$
$$\mathbb{P}[\text{success after 4 attempts}] = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = 0.9375.$$

So if we try four attempts, we will have a 93.75% probability of getting a head. Thus, four attempts is the answer.

The MATLAB / Python codes we used to generate **Figure 1.2** are shown below.

```
% MATLAB code to generate a geometric sequence
p = 1/2;
n = 1:10;
X = p.^n;
bar(n,X,'FaceColor',[0.8, 0.2,0.2]);
```

```
# Python code to generate a geometric sequence
import numpy as np
import matplotlib.pyplot as plt
p = 1/2
n = np.arange(0,10)
X = np.power(p,n)
plt.bar(n,X)
```

This warm-up exercise has perhaps raised some of your interest in the subject. However, we will not tell you everything now. We will come back to the probability in Chapter 3 when we discuss geometric random variables. In the present section, we want to make sure you have the basic mathematical tools to calculate quantities, such as a sum of fractional numbers. For example, what if we want to calculate $\mathbb{P}$[success after 107 attempts]? Is there a systematic way of performing the calculation?

Remark. You should be aware that the 93.75% only says that the probability of achieving the goal is high. If you have a bad day, you may still need more than four attempts. Therefore, when we stated the question, we asked for 90% "on average". Sometimes you may need more attempts and sometimes fewer attempts, but on average, you have a 93.75% chance of succeeding.

1.1.1 Geometric Series

A geometric series is the sum of a finite or an infinite sequence of numbers with a constant ratio between successive terms. As we have seen in the previous example, a geometric series

appears naturally in the context of discrete events. In Chapter 3 of this book, we will use geometric series when calculating the **expectation** and **moments** of a random variable.

Definition 1.1. *Let $0 < r < 1$, a* **finite geometric sequence** *of power n is a sequence of numbers*

$$\left\{1, r, r^2, \ldots, r^n\right\}.$$

An **infinite geometric sequence** *is a sequence of numbers*

$$\left\{1, r, r^2, r^3, \ldots\right\}.$$

Theorem 1.1. *The sum of a* **finite geometric series** *of power n is*

$$\sum_{k=0}^{n} r^k = 1 + r + r^2 + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r}. \tag{1.1}$$

Proof. We multiply both sides by $1 - r$. The left hand side becomes

$$\begin{aligned}\left(\sum_{k=0}^{n} r^k\right)(1 - r) &= \left(1 + r + r^2 + \cdots + r^n\right)(1 - r) \\ &= \left(1 + r + r^2 + \cdots + r^n\right) - \left(r + r^2 + r^3 + \cdots + r^{n+1}\right) \\ &\overset{(a)}{=} 1 - r^{n+1},\end{aligned}$$

where (a) holds because terms are canceled due to subtractions.

□

A corollary of Equation (1.1) is the sum of an infinite geometric sequence.

Corollary 1.1. *Let $0 < r < 1$. The sum of an* **infinite geometric series** *is*

$$\sum_{k=0}^{\infty} r^k = 1 + r + r^2 + \cdots = \frac{1}{1 - r}. \tag{1.2}$$

Proof. We take the limit in Equation (1.1). This yields

$$\sum_{k=0}^{\infty} r^k = \lim_{n\to\infty} \sum_{k=0}^{n} r^k = \lim_{n\to\infty} \frac{1 - r^{n+1}}{1 - r} = \frac{1}{1 - r}.$$

□

Remark. Note that the condition $0 < r < 1$ is important. If $r > 1$, then the limit $\lim_{n\to\infty} r^{n+1}$ in Equation (1.2) will diverge. The constant r cannot equal to 1, for otherwise the fraction $(1 - r^{n+1})/(1 - r)$ is undefined. We are not interested in the case when $r = 0$, because the sum is trivially 1: $\sum_{k=0}^{\infty} 0^k = 1 + 0^1 + 0^2 + \cdots = 1$.

Practice Exercise 1.1. Compute the infinite series $\sum_{k=2}^{\infty} \frac{1}{2^k}$.

Solution.

$$\begin{aligned}\sum_{k=2}^{\infty} \frac{1}{2^k} &= \frac{1}{4} + \frac{1}{8} + \cdots + \\ &= \frac{1}{4}\left(1 + \frac{1}{2} + \frac{1}{4} + \cdots\right) \\ &= \frac{1}{4} \cdot \frac{1}{1 - \frac{1}{2}} = \frac{1}{2}.\end{aligned}$$

Remark. You should not be confused about a geometric series and a **harmonic series**. A harmonic series concerns with the sum of $\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\}$. It turns out that[1]

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots = \infty.$$

On the other hand, a squared harmonic series $\{1, \frac{1}{2^2}, \frac{1}{3^2}, \frac{1}{4^2}, \ldots\}$ converges:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6}.$$

The latter result is known as the **Basel problem**.

We can extend the main theorem by considering more complicated series, for example the following one.

Corollary 1.2. *Let $0 < r < 1$. It holds that*

$$\sum_{k=1}^{\infty} k r^{k-1} = 1 + 2r + 3r^2 + \cdots = \frac{1}{(1-r)^2}. \tag{1.3}$$

Proof. Take the derivative on both sides of Equation (1.2). The left hand side becomes

$$\frac{d}{dr} \sum_{k=0}^{\infty} r^k = \frac{d}{dr}\left(1 + r + r^2 + \cdots\right) = 1 + 2r + 3r^2 + \cdots = \sum_{k=1}^{\infty} k r^{k-1}$$

The right hand side becomes $\dfrac{d}{dr}\left(\dfrac{1}{1-r}\right) = \dfrac{1}{(1-r)^2}$.

□

Practice Exercise 1.2. Compute the infinite sum $\sum_{k=1}^{\infty} k \cdot \frac{1}{3^k}$.

[1]This result can be found in Tom Apostol, *Mathematical Analysis*, 2nd Edition, Theorem 8.11.

Solution. We can use the derivative result:

$$\sum_{k=1}^{\infty} k \cdot \frac{1}{3^k} = 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{9} + 3 \cdot \frac{1}{27} + \cdots$$
$$= \frac{1}{3} \cdot \left(1 + 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{9} + \cdots\right) = \frac{1}{3} \cdot \frac{1}{(1-\frac{1}{3})^2} = \frac{1}{3} \cdot \frac{1}{\frac{4}{9}} = \frac{3}{4}.$$

1.1.2 Binomial Series

A geometric series is useful when handling situations such as $N-1$ failures followed by a success. However, we can easily twist the problem by asking: What is the probability of getting one head out of 3 independent coin tosses? In this case, the probability can be determined by enumerating all possible cases:

$$\begin{aligned}\mathbb{P}[\text{1 head in 3 coins}] &= \mathbb{P}[\text{H,T,T}] + \mathbb{P}[\text{T,H,T}] + \mathbb{P}[\text{T,T,H}] \\ &= \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) \\ &= \frac{3}{8}.\end{aligned}$$

Figure 1.3 illustrates the situation.

Figure 1.3: When flipping three coins independently, the probability of getting exactly one head can come from three different possibilities.

What lessons have we learned in this example? Notice that you need to enumerate all possible combinations of one head and two tails to solve this problem. The number is 3 in our example. In general, the number of combinations can be systematically studied using **combinatorics**, which we will discuss later in the chapter. However, the number of combinations motivates us to discuss another background technique known as the binomial series. The binomial series is instrumental in algebra when handling polynomials such as $(a+b)^2$ or $(1+x)^3$. It provides a valuable formula when computing these powers.

Theorem 1.2 (**Binomial theorem**). *For any real numbers a and b, the binomial series of power n is*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k, \tag{1.4}$$

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

The **binomial theorem** is valid for any real numbers a and b. The quantity $\binom{n}{k}$ reads as "n choose k". Its definition is

$$\binom{n}{k} \stackrel{\text{def}}{=} \frac{n!}{k!(n-k)!},$$

where $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$. We shall discuss the physical meaning of $\binom{n}{k}$ in Section 1.5. But we can quickly plug in the "n choose k" into the coin flipping example by letting $n = 3$ and $k = 1$:

$$\text{Number of combinations for 1 head and 2 tails} = \binom{3}{1} = \frac{3!}{1!2!} = 3.$$

So you can see why we want you to spend your precious time learning about the binomial theorem. In MATLAB and Python, $\binom{n}{k}$ can be computed using the commands as follows.

```
% MATLAB code to compute (N choose K) and K!
n = 10;
k = 2;
nchoosek(n,k)
factorial(k)
```

```
# Python code to compute (N choose K) and K!
from scipy.special import comb, factorial
n = 10
k = 2
comb(n, k)
factorial(k)
```

The binomial theorem makes the most sense when we also learn about the **Pascal's identity**.

Theorem 1.3 (**Pascal's identity**). *Let n and k be positive integers such that $k \leq n$. Then,*

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}. \tag{1.5}$$

Proof. We start by recalling the definition of $\binom{n}{k}$. This gives us

$$\begin{aligned}\binom{n}{k} + \binom{n}{k-1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-(k-1))!} \\ &= n!\left(\frac{1}{k!(n-k)!} + \frac{1}{(k-1)!(n-k+1)!}\right),\end{aligned}$$

where we factor out $n!$ to obtain the second equation. Next, we observe that

$$\begin{aligned}\frac{1}{k!(n-k)!} \times \frac{(n-k+1)}{(n-k+1)} &= \frac{n-k+1}{k!(n-k+1)!}, \\ \frac{1}{(k-1)!(n-k+1)!} \times \frac{k}{k} &= \frac{k}{k!(n-k+1)!}.\end{aligned}$$

Substituting into the previous equation we obtain

$$\begin{aligned}\binom{n}{k}+\binom{n}{k-1} &= n!\left(\frac{n-k+1}{k!(n-k+1)!}+\frac{k}{k!(n-k+1)!}\right)\\ &= n!\left(\frac{n+1}{k!(n-k+1)!}\right)\\ &= \frac{(n+1)!}{k!(n+1-k)!}\\ &= \binom{n+1}{k}.\end{aligned}$$

□

The Pascal triangle is a visualization of the coefficients of $(a+b)^n$ as shown in **Figure 1.4**. For example, when $n=5$, we know that $\binom{5}{3}=10$. However, by Pascal's identity, we know that $\binom{5}{3}=\binom{4}{2}+\binom{4}{3}$. So the number 10 is actually obtained by summing the numbers 4 and 6 of the previous row.

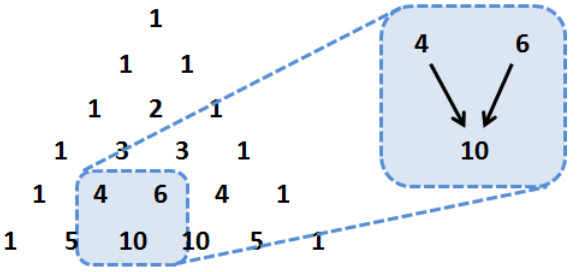


Figure 1.4: Pascal triangle for $n=0,\ldots,5$. Note that a number in one row is obtained by summing two numbers directly above it.

Practice Exercise 1.3. Find $(1+x)^3$.

Solution. Using the binomial theorem, we can show that

$$\begin{aligned}(1+x)^3 &= \sum_{k=0}^{n}\binom{3}{k}1^{3-k}x^k\\ &= 1+3x+3x^2+x^3.\end{aligned}$$

Practice Exercise 1.4. Let $0<p<1$. Find

$$\sum_{k=0}^{n}\binom{n}{k}p^{n-k}(1-p)^k.$$

Solution. By using the binomial theorem, we have

$$\sum_{k=0}^{n} \binom{n}{k} p^{n-k}(1-p)^k = (p + (1-p))^n = 1.$$

This result will be helpful when evaluating binomial random variables in Chapter 3.

We now prove the binomial theorem. Please feel free to skip the proof if this is your first time reading the book.

Proof of the binomial theorem. We prove by induction. When $n = 1$,

$$\begin{aligned}(a+b)^1 &= a + b \\ &= \sum_{k=0}^{1} a^{1-k} b^k.\end{aligned}$$

Therefore, the base case is verified. Assume up to case n. We need to verify case $n+1$.

$$\begin{aligned}(a+b)^{n+1} &= (a+b)(a+b)^n \\ &= (a+b) \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k \\ &= \sum_{k=0}^{n} \binom{n}{k} a^{n-k+1} b^k + \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^{k+1}.\end{aligned}$$

We want to apply the Pascal's identity to combine the two terms. In order to do so, we note that the second term in this sum can be rewritten as

$$\begin{aligned}\sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^{k+1} &= \sum_{k=0}^{n} \binom{n}{k} a^{n+1-k-1} b^{k+1} \\ &= \sum_{\ell=1}^{n+1} \binom{n}{\ell-1} a^{n+1-\ell} b^{\ell}, \qquad \text{where} \quad \ell = k+1 \\ &= \sum_{\ell=1}^{n} \binom{n}{\ell-1} a^{n+1-\ell} b^{\ell} + b^{n+1}.\end{aligned}$$

The first term in the sum can be written as

$$\sum_{k=0}^{n} \binom{n}{k} a^{n-k+1} b^k = \sum_{\ell=1}^{n} \binom{n}{\ell} a^{n+1-\ell} b^{\ell} + a^{n+1}, \qquad \text{where} \quad \ell = k.$$

Therefore, the two terms can be combined using Pascal's identity to yield

$$\begin{aligned}(a+b)^{n+1} &= \sum_{\ell=1}^{n} \left[\binom{n}{\ell} + \binom{n}{\ell-1} \right] a^{n+1-\ell} b^{\ell} + a^{n+1} + b^{n+1} \\ &= \sum_{\ell=1}^{n} \binom{n+1}{\ell} a^{n+1-\ell} b^{\ell} + a^{n+1} + b^{n+1} = \sum_{\ell=0}^{n+1} \binom{n+1}{\ell} a^{n+1-\ell} b^{\ell}.\end{aligned}$$

Hence, the $(n+1)$th case is also verified. By the principle of mathematical induction, we have completed the proof.

□

The end of the proof. Please join us again.

1.2 Approximation

Consider a function $f(x) = \log(1+x)$, for $x > 0$ as shown in **Figure 1.5**. This is a nonlinear function, and we all know that nonlinear functions are not fun to deal with. For example, if you want to integrate the function $\int_a^b x \log(1+x)\, dx$, then the logarithm will force you to do integration by parts. However, in many practical problems, you may not need the full range of $x > 0$. Suppose that you are only interested in values $x \ll 1$. Then the logarithm can be approximated, and thus the integral can also be approximated.

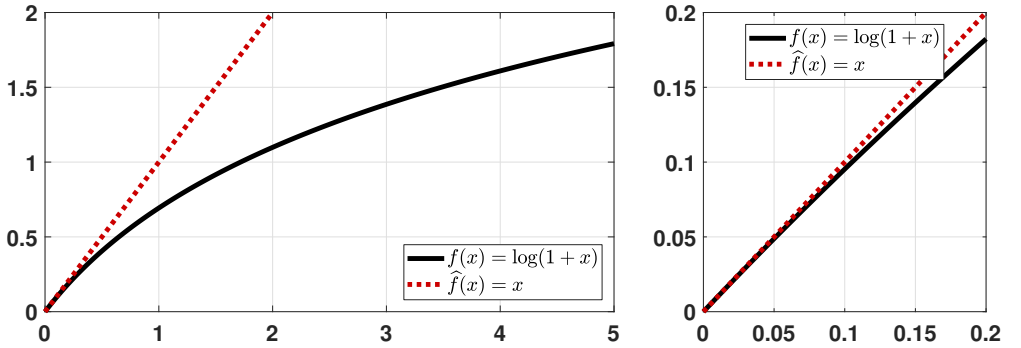


Figure 1.5: The function $f(x) = \log(1+x)$ and the approximation $\widehat{f}(x) = x$.

To see how this is even possible, we show in **Figure 1.5** the nonlinear function $f(x) = \log(1+x)$ and an approximation $\widehat{f}(x) = x$. The approximation is carefully chosen such that for $x \ll 1$, the approximation $\widehat{f}(x)$ is close to the true function $f(x)$. Therefore, we can argue that for $x \ll 1$,

$$\log(1+x) \approx x, \tag{1.6}$$

thereby simplifying the calculation. For example, if you want to integrate $x \log(1+x)$ for $0 < x < 0.1$, then the integral can be approximated by $\int_0^{0.1} x \log(1+x)\, dx \approx \int_0^{0.1} x^2\, dx = \frac{x^3}{3} = 3.33 \times 10^{-4}$. (The actual integral is 3.21×10^{-4}.) In this section we will learn about the basic approximation techniques. We will use them when we discuss limit theorems in Chapter 6, as well as various distributions, such as from binomial to Poisson.

1.2.1 Taylor approximation

Given a function $f : \mathbb{R} \to \mathbb{R}$, it is often useful to analyze its behavior by approximating f using its local information. **Taylor approximation** (or Taylor series) is one of the tools for such a task. We will use the Taylor approximation on many occasions.

Definition 1.2 (**Taylor Approximation**). *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function with infinite derivatives. Let $a \in \mathbb{R}$ be a fixed constant. The Taylor approximation of f at $x = a$ is*

$$\begin{aligned} f(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n, \end{aligned} \tag{1.7}$$

where $f^{(n)}$ denotes the nth-order derivative of f.

Taylor approximation is a geometry-based approximation. It approximates the function according to the offset, slope, curvature, and so on. According to Definition 1.2, the Taylor series has an infinite number of terms. If we use a finite number of terms, we obtain the nth-order Taylor approximation:

$$\text{First-Order :} \quad f(x) = \underbrace{f(a)}_{\text{offset}} + \underbrace{f'(a)(x-a)}_{\text{slope}} + \mathcal{O}((x-a)^2)$$

$$\text{Second-Order :} \quad f(x) = \underbrace{f(a)}_{\text{offset}} + \underbrace{f'(a)(x-a)}_{\text{slope}} + \underbrace{\frac{f''(a)}{2!}(x-a)^2}_{\text{curvature}} + \mathcal{O}((x-a)^3).$$

Here, the big-O notation $\mathcal{O}(\varepsilon^k)$ means any term that has an order at least power k. For small ε, i.e., $\varepsilon \ll 1$, a high-order term $\mathcal{O}(\varepsilon^k) \approx 0$ for large k.

Example 1.1. Let $f(x) = \sin x$. Then the Taylor approximation at $x = 0$ is

$$\begin{aligned} f(x) &\approx f(0) + f'(0)(x-0) + \frac{f''(0)}{2!}(x-0)^2 + \frac{f'''(0)}{3!}(x-0)^3 \\ &= \sin(0) + (\cos 0)(x-0) - \frac{\sin(0)}{2!}(x-0)^2 - \frac{\cos(0)}{3!}(x-0)^3 \\ &= 0 + x - 0 - \frac{x^3}{6} = x - \frac{x^3}{6}. \end{aligned}$$

We can expand further to higher orders, which yields

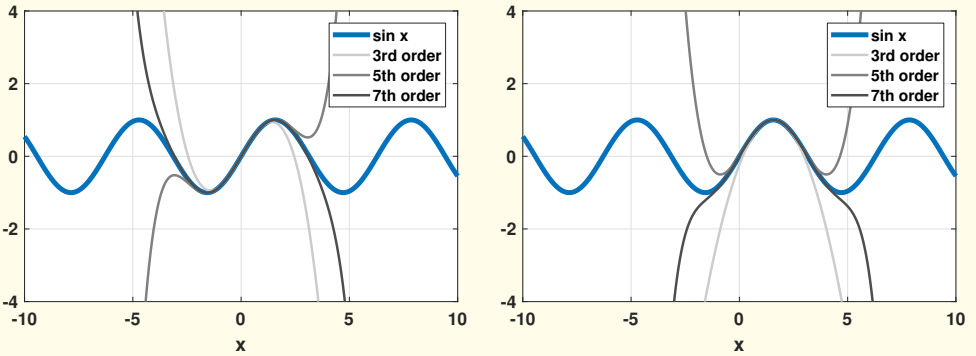
$$f(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

We show the first few approximations in **Figure 1.6**.

One should be reminded that Taylor approximation approximates a function $f(x)$ at a particular point $x = a$. Therefore, the approximation of f near $x = 0$ and the

approximation of f near $x = \pi/2$ are different. For example, the Taylor approximation at $x = \pi/2$ for $f(x) = \sin x$ is

$$\begin{aligned} f(x) &= \sin\frac{\pi}{2} + \cos\frac{\pi}{2}\left(x - \frac{\pi}{2}\right) - \frac{\sin\frac{\pi}{2}}{2!}\left(x - \frac{\pi}{2}\right)^2 - \frac{\cos\frac{\pi}{2}}{3!}\left(x - \frac{\pi}{2}\right)^3 \\ &= 1 + 0 - \frac{1}{4}\left(x - \frac{\pi}{2}\right)^2 - 0 = 1 - \frac{1}{4}\left(x - \frac{\pi}{2}\right)^2. \end{aligned}$$



(a) Approximate at $x = 0$ (b) Approximate at $x = \pi/2$

Figure 1.6: Taylor approximation of the function $f(x) = \sin x$.

1.2.2 Exponential series

An immediate application of the Taylor approximation is to derive the **exponential series**.

Theorem 1.4. *Let x be any real number. Then,*

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}. \tag{1.8}$$

Proof. Let $f(x) = e^x$ for any x. Then, the Taylor approximation around $x = 0$ is

$$\begin{aligned} f(x) &= f(0) + f'(0)(x-0) + \frac{f''(0)}{2!}(x-0)^2 + \cdots \\ &= e^0 + e^0(x-0) + \frac{e^0}{2!}(x-0)^2 + \cdots \\ &= 1 + x + \frac{x^2}{2} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}. \end{aligned}$$

□

Practice Exercise 1.5. Evaluate $\displaystyle\sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!}$.

Solution.

$$\sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

This result will be useful for **Poisson random variables** in Chapter 3.

If we substitute $x = j\theta$ where $j = \sqrt{-1}$, then we can show that

$$\begin{aligned} \underbrace{e^{j\theta}}_{=\cos\theta + j\sin\theta} &= 1 + j\theta + \frac{(j\theta)^2}{2!} + \cdots \\ &= \underbrace{\left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} + \cdots\right)}_{\text{real}} + j\underbrace{\left(\theta - \frac{\theta^3}{3!} + \cdots\right)}_{\text{imaginary}} \end{aligned}$$

Matching the real and the imaginary terms, we can show that

$$\begin{aligned} \cos\theta &= 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} + \cdots \\ \sin\theta &= \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} + \cdots \end{aligned}$$

This gives the infinite series representations of the two trigonometric functions.

1.2.3 Logarithmic approximation

Taylor approximation also allows us to find approximations to logarithmic functions. We start by presenting a lemma.

Lemma 1.1. *Let $0 < x < 1$ be a constant. Then,*

$$\log(1 + x) = x - x^2 + \mathcal{O}(x^3). \tag{1.9}$$

Proof. Let $f(x) = \log(1 + x)$. Then, the derivatives of f are

$$f'(x) = \frac{1}{(1 + x)}, \quad \text{and} \quad f''(x) = -\frac{1}{(1 + x)^2}.$$

Taylor approximation at $x = 0$ gives

$$\begin{aligned} f(x) &= f(0) + f'(0)(x - 0) + \frac{f''(0)}{2}(x - 0)^2 + \mathcal{O}(x^3) \\ &= \log 1 + \left(\frac{1}{(1 + 0)}\right) x - \left(\frac{1}{(1 + 0)^2}\right) x^2 + \mathcal{O}(x^3) \\ &= x - x^2 + \mathcal{O}(x^3). \end{aligned}$$

□

The difference between this result and the result we showed in the beginning of this section is the order of polynomials we used to approximate the logarithm:

- First-order: $\log(1+x) = x$
- Second-order: $\log(1+x) = x - x^2$.

What order of approximation is good? It depends on *where* you want the approximation to be good, and how *far* you want the approximation to go. The difference between first-order and second-order approximations is shown in **Figure 1.7**.

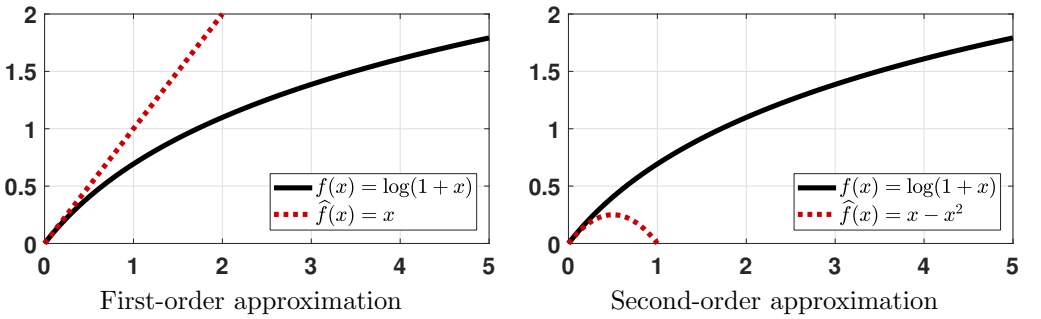


Figure 1.7: The function $f(x) = \log(1+x)$, the first-order approximation $\widehat{f}(x) = x$, and the second-order approximation $\widehat{f}(x) = x - x^2$.

Example 1.2. When we prove the **Central Limit Theorem** in Chapter 6, we need to use the following result.

$$\lim_{N\to\infty} \left(1 + \frac{s^2}{2N}\right)^N = e^{s^2/2}.$$

The proof of this equation can be done using the Taylor approximation. Consider $N \log\left(1 + \frac{s^2}{N}\right)$. By the logarithmic lemma, we can obtain the second-order approximation:

$$\log\left(1 + \frac{s^2}{2N}\right) = \frac{s^2}{2N} - \frac{s^4}{4N^2}.$$

Therefore, multiplying both sides by N yields

$$N \log\left(1 + \frac{s^2}{2N}\right) = \frac{s^2}{2} - \frac{s^4}{4N}.$$

Putting the limit $N \to \infty$ we can show that

$$\lim_{N\to\infty} \left\{N \log\left(1 + \frac{s^2}{2N}\right)\right\} = \frac{s^2}{2}.$$

Taking exponential on both sides yields

$$\exp\left\{\lim_{N\to\infty} N \log\left(1 + \frac{s^2}{2N}\right)\right\} = \exp\left\{\frac{s^2}{2}\right\}.$$

Moving the limit outside the exponential yields the result. **Figure 1.8** provides a pictorial illustration.

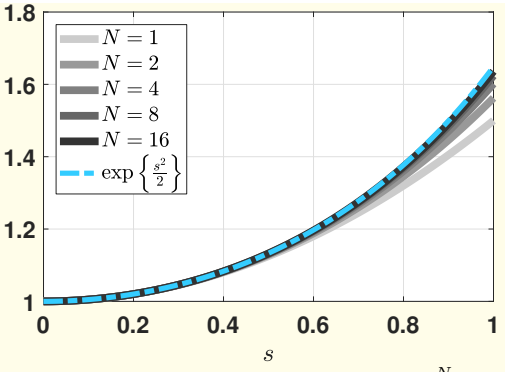


Figure 1.8: We plot a sequence of function $f_N(x) = \left(1 + \frac{s^2}{2N}\right)^N$ and its limit $f(x) = e^{s^2/2}$.

1.3 Integration

When you learned calculus, your teacher probably told you that there are two ways to compute an integral:

- **Substitution**:
$$\int f(ax)\, dx = \frac{1}{a} \int f(u)\, du.$$
- **By parts**:
$$\int u\, dv = u\, v - \int v\, du.$$

Besides these two, we want to teach you two more. The first technique is even and odd functions when integrating a function symmetrically about the y-axis. If a function is even, you just need to integrate half of the function. If a function is odd, you will get a zero. The second technique is to leverage the fact that a probability density function integrates to 1. We will discuss the first technique here and defer the second technique to Chapter 4.

Besides the two integration techniques, we will review the fundamental theorem of calculus. We will need it when we study cumulative distribution functions in Chapter 4.

1.3.1 Odd and even functions

Definition 1.3. *A function $f : \mathbb{R} \to \mathbb{R}$ is* **even** *if for any $x \in \mathbb{R}$,*

$$f(x) = f(-x), \tag{1.10}$$

and f is **odd** *if*

$$f(x) = -f(-x). \tag{1.11}$$

Essentially, an even function flips over about the y-axis, whereas an odd function flips over both the x- and y-axes.

Example 1.3. The function $f(x) = x^2 - 0.4x^4$ is even, because

$$f(-x) = (-x)^2 - 0.4(-x)^4 = x^2 - 0.4x^4 = f(x).$$

See **Figure 1.9**(a) for illustration. When integrating the function, we have

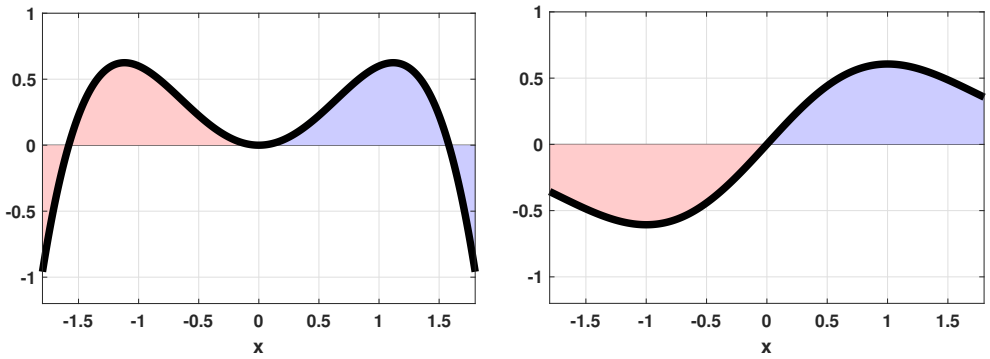
$$\int_{-1}^{1} f(x)\,dx = 2\int_{0}^{1} f(x)\,dx = 2\int_{0}^{1} x^2 - 0.4^4\,dx = 2\left[\frac{x^3}{3} - \frac{0.4}{5}x^5\right]_{x=0}^{x=1} = \frac{38}{75}.$$

Example 1.4. The function $f(x) = x\exp(-x^2/2)$ is odd, because

$$f(-x) = (-x)\exp\left\{-\frac{(-x)^2}{2}\right\} = -x\exp\left\{-\frac{x^2}{2}\right\} = -f(x).$$

See **Figure 1.9**(b) for illustration. When integrating the function, we can let $u = -x$. Then, the integral becomes

$$\begin{aligned}\int_{-1}^{1} f(x)\,dx &= \int_{-1}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx \\ &= \int_{0}^{1} f(-u)\,du + \int_{0}^{1} f(x)\,dx \\ &= -\int_{0}^{1} f(u)\,du + \int_{0}^{1} f(x)\,dx = 0.\end{aligned}$$



(a) Even function (b) Odd function

Figure 1.9: An even function is symmetric about the y-axis, and so the integration $\int_{-a}^{a} f(x)\,dx = 2\int_{0}^{a} f(x)\,dx$. An odd function is anti-symmetric about the y-axis. Thus, $\int_{-a}^{a} f(x)\,dx = 0$.

1.3.2 Fundamental Theorem of Calculus

Our following result is the **Fundamental Theorem of Calculus**. It is a handy tool that links integration and differentiation.

Theorem 1.5 (**Fundamental Theorem of Calculus**). *Let $f : [a, b] \to \mathbb{R}$ be a continuous function defined on a closed interval $[a, b]$. Then, for any $x \in (a, b)$,*

$$f(x) = \frac{d}{dx} \int_a^x f(t)\, dt, \tag{1.12}$$

Before we prove the result, let us understand the theorem if you have forgotten its meaning.

Example 1.5. Consider a function $f(t) = t^2$. If we integrate the function from 0 to x, we will obtain another function

$$F(x) \stackrel{\text{def}}{=} \int_0^x f(t)\, dt = \int_0^x t^2\, dt = \frac{x^3}{3}.$$

On the other hand, we can differentiate $F(x)$ to obtain $f(x)$:

$$f(x) = \frac{d}{dx} F(x) = \frac{d}{dx} \frac{x^3}{3} = x^2.$$

The fundamental theorem of calculus basically puts the two together:

$$f(x) = \frac{d}{dx} \int_0^x f(t)\, dt.$$

That's it. Nothing more and nothing less.

How can the fundamental theorem of calculus ever be useful when studying probability? Very soon you will learn two concepts: **probability density function** and **cumulative distribution function**. These two functions are related to each other by the fundamental theorem of calculus. To give you a concrete example, we write down the probability density function of an exponential random variable. (Please do not panic about the exponential random variable. Just think of it as a "rapidly decaying" function.)

$$f(x) = e^{-x}, \quad x \geq 0.$$

It turns out that the cumulative distribution function is

$$F(x) = \int_0^x f(t)\, dt = \int_0^x e^{-t}\, dt = 1 - e^{-x}.$$

You can also check that $f(x) = \frac{d}{dx} F(x)$. The fundamental theorem of calculus says that if you tell me $F(x) = \int_0^x e^{-t}\, dt$ (for whatever reason), I will be able to tell you that $f(x) = e^{-x}$ merely by visually inspecting the integrand without doing the differentiation.

Figure 1.10 illustrates the pair of functions $f(x) = e^{-x}$ and $F(x) = 1 - e^{-x}$. One thing you should notice is that the *height* of $F(x)$ is the area under the curve of $f(t)$ from $-\infty$ to x. For example, in **Figure 1.10** we show the area under the curve from 0 to 2. Correspondingly in $F(x)$, the height is $F(2)$.

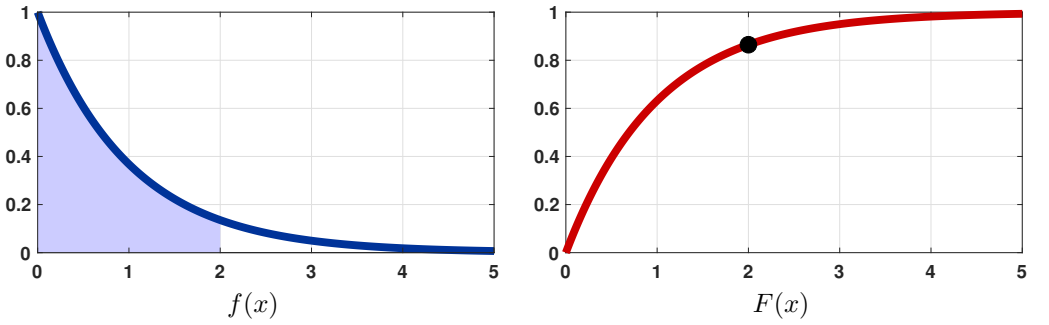


Figure 1.10: The pair of functions $f(x) = e^{-x}$ and $F(x) = 1 - e^{-x}$

The following proof of the Fundamental Theorem of Calculus can be skipped if it is your first time reading the book.

Proof. Our proof is based on Stewart (6th Edition), Section 5.3. Define the integral as a function F:

$$F(x) = \int_a^x f(t)\, dt.$$

The derivative of F with respect to x is

$$\begin{aligned}
\frac{d}{dx}F(x) &= \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} \\
&= \lim_{h\to 0} \frac{1}{h}\left(\int_a^{x+h} f(t)\, dt - \int_a^x f(t)\, dt \right) \\
&= \lim_{h\to 0} \frac{1}{h} \int_x^{x+h} f(t)\, dt \\
&\overset{(a)}{\leq} \lim_{h\to 0} \frac{1}{h} \int_x^{x+h} \left\{ \max_{x\leq\tau\leq x+h} f(\tau) \right\} dt \\
&= \lim_{h\to 0} \left\{ \max_{x\leq\tau\leq x+h} f(\tau) \right\}.
\end{aligned}$$

Here, the inequality in (a) holds because

$$f(t) \leq \max_{x\leq\tau\leq x+h} f(\tau)$$

for all $x \leq t \leq x + h$. The maximum exists because f is continuous in a closed interval.

Using the parallel argument, we can show that

$$\begin{aligned}\frac{d}{dx}F(x) &= \lim_{h\to 0}\frac{F(x+h)-F(x)}{h}\\ &= \lim_{h\to 0}\frac{1}{h}\left(\int_a^{x+h} f(t)\,dt - \int_a^{x} f(t)\,dt\right)\\ &= \lim_{h\to 0}\frac{1}{h}\int_x^{x+h} f(t)\,dt\\ &\geq \lim_{h\to 0}\frac{1}{h}\int_x^{x+h}\left\{\min_{x\leq\tau\leq x+h} f(\tau)\right\}\,dt\\ &= \lim_{h\to 0}\left\{\min_{x\leq\tau\leq x+h} f(\tau)\right\}.\end{aligned}$$

Combining the two results, we have that

$$\lim_{h\to 0}\left\{\min_{x\leq\tau\leq x+h} f(\tau)\right\} \leq \frac{d}{dx}F(x) \leq \lim_{h\to 0}\left\{\max_{x\leq\tau\leq x+h} f(\tau)\right\}.$$

However, since the two limits are both converging to $f(x)$ as $h\to 0$, we conclude that $\frac{d}{dx}F(x) = f(x)$.

□

Remark. An alternative proof is to use Mean Value Theorem in terms of Riemann-Stieltjes integrals (see, e.g., Tom Apostol, *Mathematical Analysis*, 2nd edition, Theorem 7.34). To handle more general functions such as delta functions, one can use techniques in Lebesgue's integration. However, this is beyond the scope of this book.

This is the end of the proof. Please join us again.

In many practical problems, the fundamental theorem of calculus needs to be used in conjunction with the **chain rule**.

Corollary 1.3. *Let $f:[a,b]\to\mathbb{R}$ be a continuous function defined on a closed interval $[a,b]$. Let $g:\mathbb{R}\to[a,b]$ be a continuously differentiable function. Then, for any $x\in(a,b)$,*

$$\frac{d}{dx}\int_a^{g(x)} f(t)\,dt = g'(x)\cdot f(g(x)). \tag{1.13}$$

Proof. We can prove this with the chain rule: Let $y=g(x)$. Then we have

$$\frac{d}{dx}\int_a^{g(x)} f(t)\,dt = \frac{dy}{dx}\cdot\frac{d}{dy}\int_a^{y} f(t)\,dt = g'(x)\,f(y),$$

which completes the proof.

□

Practice Exercise 1.6. Evaluate the integral

$$\frac{d}{dx}\int_0^{x-\mu}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{t^2}{2\sigma^2}\right\}\,dt.$$

Solution. Let $y = x - \mu$. Then by using the fundamental theorem of calculus, we can show that

$$\begin{aligned}\frac{d}{dx}\int_0^{x-\mu}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{t^2}{2\sigma^2}\right\}\,dt &= \frac{dy}{dx}\cdot\frac{d}{dy}\int_0^{y}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{t^2}{2\sigma^2}\right\}\,dt\\ &= \frac{d(x-\mu)}{dx}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{y^2}{2\sigma^2}\right\}\\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.\end{aligned}$$

This result will be useful when we do linear transformations of a Gaussian random variable in Chapter 4.

1.4 Linear Algebra

The two most important subjects for data science are *probability*, which is the subject of the book you are reading, and *linear algebra*, which concerns matrices and vectors. We cannot cover linear algebra in detail because this would require another book. However, we need to highlight some ideas that are important for doing data analysis.

1.4.1 Why do we need linear algebra in data science?

Consider a dataset of the crime rate of several cities as shown below, downloaded from `https://web.stanford.edu/~hastie/StatLearnSparsity/data.html`.

The table shows that the crime rate depends on several factors such as funding for the police department, the percentage of high school graduates, etc.

city	`crime rate`	`funding`	`hs`	`no-hs`	`college`	`college4`
1	478	40	74	11	31	20
2	494	32	72	11	43	18
3	643	57	71	18	16	16
4	341	31	71	11	25	19
⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	940	66	67	26	18	16

What questions can we ask about this table? We can ask: What is the most influential cause of the crime rate? What are the leading contributions to the crime rate? To answer these questions, we need to describe these numbers. One way to do it is to put the numbers in matrices and vectors. For example,

$$\boldsymbol{y}_{\text{crime}} = \begin{bmatrix} 478 \\ 494 \\ \vdots \\ 940 \end{bmatrix}, \quad \boldsymbol{x}_{\text{fund}} = \begin{bmatrix} 40 \\ 32 \\ \vdots \\ 66 \end{bmatrix}, \quad \boldsymbol{x}_{\text{hs}} = \begin{bmatrix} 74 \\ 72 \\ \vdots \\ 67 \end{bmatrix}, \dots$$

With this vector expression of the data, the analysis questions can roughly be translated to finding β's in the following equation:

$$\boldsymbol{y}_{\text{crime}} = \beta_{\text{fund}}\boldsymbol{x}_{\text{fund}} + \beta_{\text{hs}}\boldsymbol{x}_{\text{hs}} + \cdots + \beta_{\text{college4}}\boldsymbol{x}_{\text{college4}}.$$

This equation offers a lot of useful insights. First, it is a **linear model** of $\boldsymbol{y}_{\text{crime}}$. We call it a linear model because the observable $\boldsymbol{y}_{\text{crime}}$ is written as a **linear combination** of the variables $\boldsymbol{x}_{\text{fund}}, \boldsymbol{x}_{\text{hs}}$, etc. The linear model assumes that the variables are scaled and added to generate the observed phenomena. This assumption is not always realistic, but it is often a fair assumption that greatly simplifies the problem. For example, if we can show that all β's are zero except β_{fund}, then we can conclude that the crime rate is solely dependent on the police funding. If two variables are correlated, e.g., high school graduate and college graduate, we would expect the β's to change simultaneously.

The linear model can further be simplified to a matrix-vector equation:

$$\begin{bmatrix} | \\ | \\ \boldsymbol{y}_{\text{crime}} \\ | \\ | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ | & | & & | \\ \boldsymbol{x}_{\text{fund}} & \boldsymbol{x}_{\text{hs}} & \cdots & \boldsymbol{x}_{\text{college4}} \\ | & | & & | \\ | & | & & | \end{bmatrix} \begin{bmatrix} \beta_{\text{fund}} \\ \beta_{\text{hs}} \\ \vdots \\ \beta_{\text{college4}} \end{bmatrix}$$

Here, the lines "|" emphasize that the vectors are column vectors. If we denote the matrix in the middle as $\boldsymbol{A}$ and the vector as $\boldsymbol{\beta}$, then the equation is equivalent to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\beta}$. So we can find $\boldsymbol{\beta}$ by appropriately inverting the matrix $\boldsymbol{A}$. If two columns of $\boldsymbol{A}$ are dependent, we will not be able to resolve the corresponding β's uniquely.

As you can see from the above data analysis problem, matrices and vectors offer a way to describe the data. We will discuss the calculations in Chapter 7. However, to understand how to interpret the results from the matrix-vector equations, we need to review some basic ideas about matrices and vectors.

1.4.2 Everything you need to know about linear algebra

Throughout this book, you will see different sets of notations. For linear algebra, we also have a set of notations. We denote $\boldsymbol{x} \in \mathbb{R}^d$ a d-dimensional vector taking real numbers as its entries. An M-by-N matrix is denoted as $\boldsymbol{X} \in \mathbb{R}^{M \times N}$. The transpose of a matrix is denoted as $\boldsymbol{X}^T$. A matrix $\boldsymbol{X}$ can be viewed according to its columns and its rows:

$$\boldsymbol{X} = \begin{bmatrix} | & | & & | \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_N \\ | & | & & | \end{bmatrix}, \quad \text{and} \quad \boldsymbol{X} = \begin{bmatrix} — & \boldsymbol{x}^1 & — \\ — & \boldsymbol{x}^2 & — \\ & \vdots & \\ — & \boldsymbol{x}^M & — \end{bmatrix}.$$

Here, $\boldsymbol{x}_j$ denotes the jth column of $\boldsymbol{X}$, and $\boldsymbol{x}^i$ denotes the ith row of $\boldsymbol{X}$. The (i,j)th element of $\boldsymbol{X}$ is denoted as x_{ij} or $[\boldsymbol{X}]_{ij}$. The identity matrix is denoted as $\boldsymbol{I}$. The ith column of $\boldsymbol{I}$ is denoted as $\boldsymbol{e}_i = [0, \ldots, 1, \ldots, 0]^T$, and is called the ith **standard basis vector**. An all-zero vector is denoted as $\boldsymbol{0} = [0, \ldots, 0]^T$.

What is the most important thing to know about linear algebra? From a data analysis point of view, **Figure 1.11** gives us the answer. The picture is straightforward, but it captures all the essence. In almost all the data analysis problems, ultimately, there are three things we care about: (i) The observable vector $\boldsymbol{y}$, (ii) the variable vectors $\boldsymbol{x}_n$, and (iii) the coefficients β_n. The set of variable vectors $\{\boldsymbol{x}_n\}_{n=1}^N$ **spans** a vector space in which all vectors are living. Some of these variable vectors are correlated, and some are not. However, for the sake of this discussion, let us assume they are independent of each other. Then for any observable vector $\boldsymbol{y}$, we can always project $\boldsymbol{y}$ in the directions determined by $\{\boldsymbol{x}_n\}_{n=1}^N$. The projection of $\boldsymbol{y}$ onto $\boldsymbol{x}_n$ is the coefficient β_n. A larger value of β_n means that the variable $\boldsymbol{x}_n$ has more contributions.

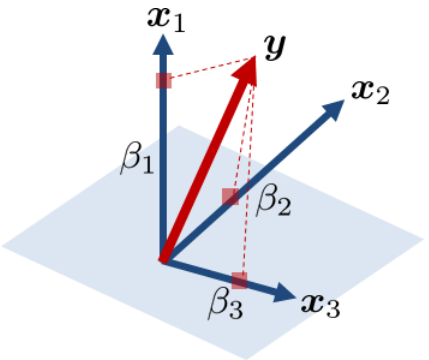


Figure 1.11: Representing an observable vector $\boldsymbol{y}$ by a linear combination of variable vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$. The combination weights are $\beta_1, \beta_2, \beta_3$.

Why is this picture so important? Because most of the data analysis problems can be expressed, or approximately expressed, by the picture:

$$\boldsymbol{y} = \sum_{n=1}^{N} \beta_n \boldsymbol{x}_n.$$

If you recall the crime rate example, this equation is precisely the linear model we used to describe the crime rate. This equation can also describe many other problems.

Example 1.6. **Polynomial fitting**. Consider a dataset of pairs of numbers (t_m, y_m) for $m = 1, \ldots, M$, as shown in **Figure 1.12**. After a visual inspection of the dataset, we propose to use a line to fit the data. A line is specified by the equation

$$y_m = at_m + b, \qquad m = 1, \ldots, M,$$

where $a \in \mathbb{R}$ is the slope and $b \in \mathbb{R}$ is the y-intercept. The goal of this problem is to find one line (which is fully characterized by (a, b)) such that it has the best fit to *all* the data pairs (t_m, y_m) for $m = 1, \ldots, M$. This problem can be described in matrices

and vectors by noting that

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{a}_{\beta_1} \underbrace{\begin{bmatrix} t_1 \\ \vdots \\ t_M \end{bmatrix}}_{\boldsymbol{x}_1} + \underbrace{b}_{\beta_2} \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{\boldsymbol{x}_2},$$

or more compactly,

$$\boldsymbol{y} = \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2.$$

Here, $\boldsymbol{x}_1 = [t_1, \ldots, t_M]^T$ contains all the variable values, and $\boldsymbol{x}_2 = [1, \ldots, 1]^T$ contains a constant offset.

t_m	y_m
0.1622	2.1227
0.7943	3.3354
⋮	⋮
0.7379	3.4054
0.2691	2.5672
0.4228	2.3796
0.6020	3.2942

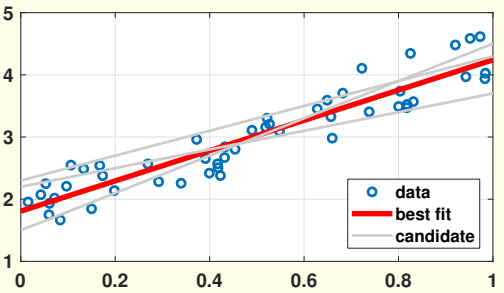


Figure 1.12: Example of fitting a set of data points. The problem can be described by $\boldsymbol{y} = \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2$.

Example 1.7. **Image compression**. The JPEG compression for images is based on the concept of **discrete cosine transform** (DCT). The DCT consists of a set of **basis vectors**, or $\{\boldsymbol{x}_n\}_{n=1}^N$ using our notation. In the most standard setting, each basis vector $\boldsymbol{x}_n$ consists of 8×8 pixels, and there are $N = 64$ of these $\boldsymbol{x}_n$'s. Given an image, we can partition the image into M small blocks of 8×8 pixels. Let us call one of these blocks $\boldsymbol{y}$. Then, DCT represents the observation $\boldsymbol{y}$ as a linear combination of the DCT basis vectors:

$$\boldsymbol{y} = \sum_{n=1}^{N} \beta_n \boldsymbol{x}_n.$$

The coefficients $\{\beta_n\}_{n=1}^N$ are called the DCT coefficients. They provide a **representation** of $\boldsymbol{y}$, because once we know $\{\beta_n\}_{n=1}^N$, we can completely describe $\boldsymbol{y}$ because the basis vectors $\{\boldsymbol{x}_n\}_{n=1}^N$ are known and fixed. The situation is depicted in **Figure 1.13**.

How can we compress images using DCT? In the 1970s, scientists found that most images have strong leading DCT coefficients but weak tail DCT coefficients. In other words, among the $N = 64$ β_n's, only the first few are important. If we truncate the number of DCT coefficients, we can effectively compress the number of bits required to represent the image.

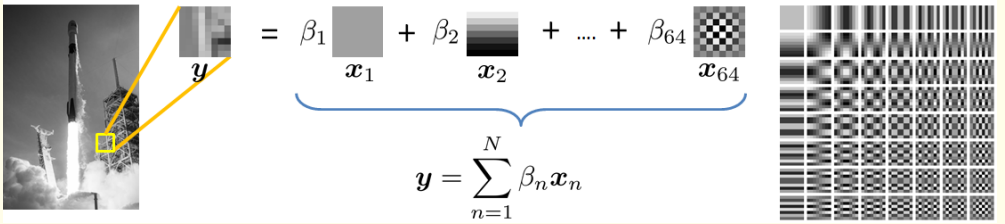


Figure 1.13: JPEG image compression is based on the concept of discrete cosine transform, which can be formulated as a matrix-vector problem.

We hope by now you are convinced of the importance of matrices and vectors in the context of data science. They are not "yet another" subject but an essential tool you must know how to use. So, what are the technical materials you must master? Here we go.

1.4.3 Inner products and norms

We assume that you know the basic operations such as matrix-vector multiplication, taking the transpose, etc. If you have forgotten these, please consult any undergraduate linear algebra textbook such as Gilbert Strang's *Linear Algebra and its Applications*. We will highlight a few of the most important operations for our purposes.

Definition 1.4 (**Inner product**). *Let* $\boldsymbol{x} = [x_1, \ldots, x_N]^T$, *and* $\boldsymbol{y} = [y_1, \ldots, y_N]^T$. *The inner product* $\boldsymbol{x}^T\boldsymbol{y}$ *is*

$$\boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{N} x_i y_i. \tag{1.14}$$

Practice Exercise 1.7. Let $\boldsymbol{x} = [1,\ 0,\ -1]^T$, and $\boldsymbol{y} = [3,\ 2,\ 0]^T$. Find $\boldsymbol{x}^T\boldsymbol{y}$.

Solution. The inner product is $\boldsymbol{x}^T\boldsymbol{y} = (1)(3) + (0)(2) + (-1)(0) = 3$.

Inner products are important because they tell us how two vectors are correlated. **Figure 1.14** depicts the geometric meaning of an inner product. If two vectors are correlated (i.e., nearly parallel), then the inner product will give us a large value. Conversely, if the two vectors are close to perpendicular, then the inner product will be small. Therefore, the inner product provides a measure of the closeness/similarity between two vectors.

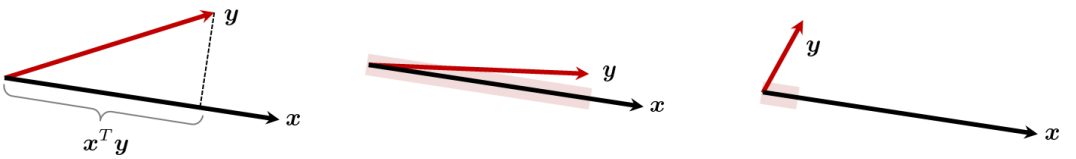


Figure 1.14: Geometric interpretation of inner product: We project one vector onto the other vector. The projected distance is the inner product.

Creating vectors and computing the inner products are straightforward in MATLAB. We simply need to define the column vectors `x` and `y` by using the command `[]` with `;` to denote the next row. The inner product is done using the transpose operation `x'` and vector multiplication `*`.

```
% MATLAB code to perform an inner product
x = [1 0 -1];
y = [3 2 0];
z = x'*y;
```

In Python, constructing a vector is done using the command `np.array`. Inside this command, one needs to enter the array. For a column vector, we write `[[1],[2],[3]]`, with an outer `[]`, and three inner `[]` for each entry. If the vector is a row vector, the one can omit the inner `[]`'s by just calling `np.array([1, 2, 3])`. Given two column vectors `x` and `y`, the inner product is computed via `np.dot(x.T,y)`, where `np.dot` is the command for inner product, and `x.T` returns the transpose of `x`. One can also call `np.transpose(x)`, which is the same as `x.T`.

```
# Python code to perform an inner product
import numpy as np
x = np.array([[1],[0],[-1]])
y = np.array([[3],[2],[0]])
z = np.dot(np.transpose(x),y)
print(z)
```

In data analytics, the inner product of two vectors can be useful. Consider the vectors in **Table 1.1**. Just from looking at the numbers, you probably will not see anything wrong. However, let's compute the inner products. It turns out that $\boldsymbol{x}_1^T\boldsymbol{x}_2 = -0.0031$, whereas $\boldsymbol{x}_1^T\boldsymbol{x}_3 = 2.0020$. There is almost no correlation between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, but there is a substantial correlation between $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$. What happened? The vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are random vectors constructed independently and uncorrelated to each other. The last vector $\boldsymbol{x}_3$ was constructed by $\boldsymbol{x}_3 = 2\boldsymbol{x}_1 - \pi/1000$. Since $\boldsymbol{x}_3$ is completely constructed from $\boldsymbol{x}_1$, they have to be correlated.

$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$
0.0006	−0.0011	−0.0020
−0.0014	−0.0024	−0.0059
−0.0034	0.0073	−0.0099
⋮	⋮	⋮
0.0001	−0.0066	−0.0030
0.0074	0.0046	0.0116
0.0007	−0.0061	−0.0017

Table 1.1: Three example vectors.

One caveat for this example is that the naive inner product $\boldsymbol{x}_i^T\boldsymbol{x}_j$ is scale-dependent. For example, the vectors $\boldsymbol{x}_3 = \boldsymbol{x}_1$ and $\boldsymbol{x}_3 = 1000\boldsymbol{x}_1$ have the same amount of correlation,

but the simple inner product will give a larger value for the latter case. To solve this problem we first define the **norm** of the vectors:

Definition 1.5 (**Norm**). *Let $\boldsymbol{x} = [x_1, \ldots, x_N]^T$ be a vector. The ℓ_p-norm of $\boldsymbol{x}$ is*

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{N} x_i^p \right)^{1/p}, \tag{1.15}$$

for any $p \geq 1$.

The norm essentially tells us the **length** of the vector. This is most obvious if we consider the ℓ_2-norm:

$$\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{N} x_i^2 \right)^{1/2}.$$

By taking the square on both sides, one can show that $\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^T\boldsymbol{x}$. This is called the **squared ℓ_2-norm**, and is the sum of the squares.

On MATLAB, computing the norm is done using the command `norm`. Here, we can indicate the types of norms, e.g., `norm(x,1)` returns the ℓ_1-norm whereas `norm(x,2)` returns the ℓ_2-norm (which is also the default).

```
% MATLAB code to compute the norm
x = [1 0 -1];
x_norm = norm(x);
```

On Python, the norm command is listed in the `np.linalg`. To call the ℓ_1-norm, we use `np.linalg.norm(x,1)`, and by default the ℓ_2-norm is `np.linalg.norm(x)`.

```
# Python code to compute the norm
import numpy as np
x = np.array([[1],[0],[-1]])
x_norm = np.linalg.norm(x)
```

Using the norm, one can define an angle called the **cosine angle** between two vectors.

Definition 1.6. *The* **cosine angle** *between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is*

$$\cos\theta = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2}. \tag{1.16}$$

The difference between the cosine angle and the basic inner product is the **normalization** in the denominator, which is the product $\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2$. This normalization factor scales the vector $\boldsymbol{x}$ to $\boldsymbol{x}/\|\boldsymbol{x}\|_2$ and $\boldsymbol{y}$ to $\boldsymbol{y}/\|\boldsymbol{y}\|_2$. The scaling makes the length of the new vector equal to unity, but it does not change the vector's orientation. Therefore, the cosine angle is not affected by a very long vector or a very short vector. Only the angle matters. See **Figure 1.15**.

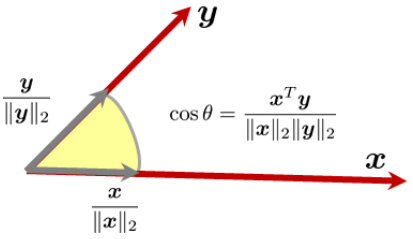


Figure 1.15: The cosine angle is the inner product divided by the norms of the vectors.

Going back to the previous example, after normalization we can show that the cosine angle between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is $\cos\theta_{1,2} = -0.0031$, whereas the cosine angle between $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$ is $\cos\theta_{1,3} = 0.8958$. There is still a strong correlation between $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$, but now using the cosine angle the value is between -1 and $+1$.

Remark 1: There are other norms one can use. The ℓ_1-norm is useful for **sparse** models where we want to have the fewest possible non-zeros. The ℓ_1-norm of $\boldsymbol{x}$ is

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{N} |x_i|,$$

which is the sum of absolute values. The ℓ_∞-norm picks the maximum of $\{x_1, \ldots, x_N\}$:

$$\begin{aligned}\|\boldsymbol{x}\|_\infty &= \lim_{p\to\infty} \left(\sum_{i=1}^{N} x_i^p\right)^{1/p} \\ &= \max\{x_1, \ldots, x_N\},\end{aligned}$$

because as $p \to \infty$, only the largest element will be amplified.

Remark 2: The standard ℓ_2-norm is a circle: Just consider $\boldsymbol{x} = [x_1, x_2]^T$. The norm is $\|\boldsymbol{x}\|_2 = \sqrt{x_1^2 + x_2^2}$. We can convert the circle to ellipses by considering a weighted norm.

Definition 1.7 (**Weighted ℓ_2-norm square**). *Let $\boldsymbol{x} = [x_1, \ldots, x_N]^T$ and let $\boldsymbol{W} = diag(w_1, \ldots, w_N)$ be a non-negative diagonal matrix. The weighted ℓ_2-norm square of $\boldsymbol{x}$ is*

$$\begin{aligned}\|\boldsymbol{x}\|_{\boldsymbol{W}}^2 &= \boldsymbol{x}^T \boldsymbol{W} \boldsymbol{x} \\ &= \begin{bmatrix} x_1 & \ldots & x_N \end{bmatrix} \begin{bmatrix} w_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & w_N \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \sum_{i=1}^{N} w_i x_i^2. \end{aligned} \tag{1.17}$$

The geometry of the weighted ℓ_2-norm is determined by the matrix $\boldsymbol{W}$. For example, if $\boldsymbol{W} = \boldsymbol{I}$ (the identity operator), then $\|\boldsymbol{x}\|_{\boldsymbol{W}}^2 = \|\boldsymbol{x}\|_2^2$, which defines a circle. If $\boldsymbol{W}$ is any "non-negative" matrix[2], then $\|\boldsymbol{x}\|_{\boldsymbol{W}}^2$ defines an ellipse.

[2]The technical term for these matrices is *positive semi-definite* matrices.

In MATLAB, the weighted inner product is just a sequence of two matrix-vector multiplications. This can be done using the command `x'*W*x` as shown below.

```
% MATLAB code to compute the weighted norm
W = [1 2 3; 4 5 6; 7 8 9];
x = [2; -1; 1];
z = x'*W*x
```

In Python, constructing the matrix $\boldsymbol{W}$ and the column vector $\boldsymbol{x}$ is done using `np.array`. The matrix-vector multiplication is done using two `np.dot` commands: one for `np.dot(W,x)` and the other one for `np.dot(x.T, np.dot(W,x))`.

```
# Python code to compute the weighted norm
import numpy as np
W = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
x = np.array([[2],[-1],[1]])
z = np.dot(x.T, np.dot(W,x))
print(z)
```

1.4.4 Matrix calculus

The last linear algebra topic we need to review is matrix calculus. As its name indicates, matrix calculus is about the differentiation of matrices and vectors. Why do we need differentiation for matrices and vectors? Because we want to find the **minimum or maximum** of a scalar function with a vector input.

Let us go back to the crime rate problem we discussed earlier. Given the data, we want to find the model coefficients $\beta_1, \ldots, \beta_N$ such that the variables can best explain the observation. In other words, we want to minimize the deviation between $\boldsymbol{y}$ and the prediction offered by our model:

$$\underset{\beta_1,\ldots,\beta_N}{\text{minimize}} \quad \left\| \boldsymbol{y} - \sum_{n=1}^{N} \beta_n \boldsymbol{x}_n \right\|^2 .$$

This equation is self-explanatory. The norm $\|\clubsuit - \heartsuit\|^2$ measures the deviation. If $\boldsymbol{y}$ can be perfectly explained by $\{\boldsymbol{x}_n\}_{n=1}^N$, then the norm can eventually go to zero by finding a good set of $\{\beta_1, \ldots, \beta_N\}$. The symbol $\underset{\beta_1,\ldots,\beta_N}{\text{minimize}}$ means to minimize the function by finding $\{\beta_1, \ldots, \beta_N\}$. Note that the norm is taking a vector as the input and generating a scalar as the output. It can be expressed as

$$\varepsilon(\boldsymbol{\beta}) \overset{\text{def}}{=} \left\| \boldsymbol{y} - \sum_{n=1}^{N} \beta_n \boldsymbol{x}_n \right\|^2 ,$$

to emphasize this relationship. Here we define $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_N]^T$ as the collection of all coefficients.

Given this setup, how would you determine $\boldsymbol{\beta}$ such that the deviation is minimized? Our calculus teachers told us that we could take the function's derivative and set it to zero

for scalar problems. It is the same story for vectors. What we do is to take the derivative of the error and set it equal to zero:

$$\frac{d}{d\boldsymbol{\beta}}\,\varepsilon(\boldsymbol{\beta}) = 0.$$

Now the question arises, how do we take the derivatives of $\varepsilon(\boldsymbol{\beta})$ when it takes a vector as input? If we can answer this question, we will find the best $\boldsymbol{\beta}$. The answer is straightforward. Since the function has one output and many inputs, take the derivative for each element independently. This is called the **scalar differentiation of vectors**.

Definition 1.8 (**Scalar differentiation of vectors**)**.** *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a differentiable scalar function, and let $y = f(\boldsymbol{x})$ for some input $\boldsymbol{x} \in \mathbb{R}^N$. Then,*

$$\frac{dy}{d\boldsymbol{x}} = \begin{bmatrix} dy/dx_1 \\ \vdots \\ dy/dx_N \end{bmatrix}.$$

As you can see from this definition, there is nothing conceptually challenging here. The only difficulty is that things can get tedious because there will be many terms. However, the good news is that mathematicians have already compiled a list of identities for common matrix differentiation. So instead of deriving every equation from scratch, we can enjoy the fruit of their hard work by referring to those formulae. The best place to find these equations is the *Matrix Cookbook* by Petersen and Pedersen.[3] Here, we will mention two of the most useful results.

Example 1.8. Let $y = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ for any matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$. Find $\frac{dy}{d\boldsymbol{x}}$.

Solution.

$$\frac{d}{d\boldsymbol{x}}\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^T\boldsymbol{x}.$$

Now, if $\boldsymbol{A}$ is symmetric, i.e., $\boldsymbol{A} = \boldsymbol{A}^T$, then

$$\frac{d}{d\boldsymbol{x}}\left(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right) = 2\boldsymbol{A}\boldsymbol{x}.$$

Example 1.9. Let $\varepsilon = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2$, where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is symmetric. Find $\frac{d\varepsilon}{d\boldsymbol{x}}$.

Solution. First, we note that

$$\varepsilon = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{y}^T\boldsymbol{y}.$$

[3] https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Taking the derivative with respect to $\boldsymbol{x}$ yields

$$\begin{aligned}\frac{d\varepsilon}{d\boldsymbol{x}} &= 2\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{A}^T\boldsymbol{y} \\ &= 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}).\end{aligned}$$

Going back to the crime rate problem, we can now show that

$$0 = \frac{d\varepsilon}{d\boldsymbol{\beta}}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}).$$

Therefore, the solution is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}.$$

As you can see, if we do not have access to the matrix calculus, we will not be able to solve the minimization problem. (There are alternative paths that do not require matrix calculus, but they require an understanding of linear subspaces and properties of the projection operators. So in some sense, matrix calculus is the easiest way to solve the problem.) When we discuss the linear regression methods in Chapter 7, we will cover the interpretation of the inverses and related topics.

In MATLAB and Python, matrix inversion is done using the command `inv` in MATLAB and `np.linalg.inv` in Python. Below is an example in Python.

```
# Python code to compute a matrix inverse
import numpy as np
X      = np.array([[1, 3], [-2, 7], [0, 1]])
XtX    = np.dot(X.T, X)
XtXinv = np.linalg.inv(XtX)
print(XtXinv)
```

Sometimes, instead of computing the matrix inverse we are more interested in solving a linear equation $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$ (the solution of which is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}$). In both MATLAB and Python, there are built-in commands to do this. In MATLAB, the command is \ (backslash).

```
% MATLAB code to solve X beta = y
X      = [1 3; -2 7; 0 1];
y      = [2; 1; 0];
beta   = X\y;
```

In Python, the built-in command is `np.linalg.lstsq`.

```
# Python code to solve X beta = y
import numpy as np
X      = np.array([[1, 3], [-2, 7], [0, 1]])
y      = np.array([[2],[1],[0]])
beta   = np.linalg.lstsq(X, y, rcond=None)[0]
print(beta)
```

Closing remark: In this section, we have given a brief introduction to a few of the most relevant concepts in linear algebra. We will introduce further concepts in linear algebra in later chapters, such as eigenvalues, principal component analysis, linear transformations, and regularization, as they become useful for our discussion.

1.5 Basic Combinatorics

The last topic we review in this chapter is **combinatorics**. Combinatorics concerns the number of configurations that can be obtained from certain discrete experiments. It is useful because it provides a systematic way of enumerating cases. Combinatorics often becomes very challenging as the complexity of the event grows. However, you may rest assured that in this book, we will not tackle the more difficult problems of combinatorics; we will confine our discussion to two of the most basic principles: **permutation** and **combination**.

1.5.1 Birthday paradox

To motivate the discussion of combinatorics, let us start with the following problem. Suppose there are 50 people in a room. What is the probability that at least one pair of people have the same birthday (month and day)? (We exclude Feb. 29 in this problem.)

The first thing you might be thinking is that since there are 365 days, we need at least 366 people to ensure that one pair has the same birthday. Therefore, the chance that 2 of 50 people have the same birthday is low. This seems reasonable, but let's do a simulated experiment. In **Figure 1.16** we plot the probability as a function of the number of people. For a room containing 50 people, the probability is 97%. To get a 50% probability, we just need 23 people! How is this possible?

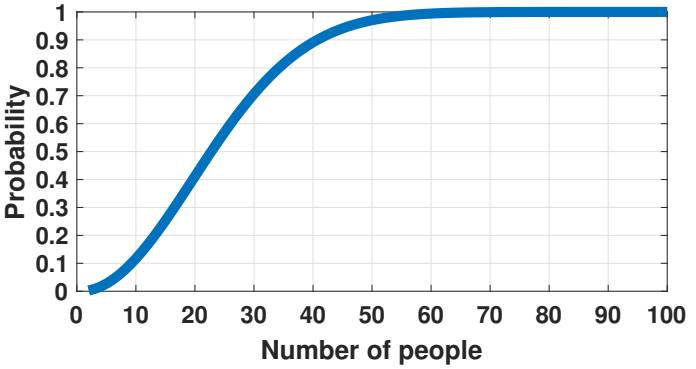


Figure 1.16: The probability for two people in a group to have the same birthday as a function of the number of people in the group.

If you think about this problem more deeply, you will probably realize that to solve the problem, we must carefully enumerate all the possible configurations. How can we do this? Well, suppose you walk into the room and sequentially pick two people. The probability

that they have *different* birthdays is

$$\mathbb{P}[\text{The first 2 people have different birthdays}] = \frac{365}{365} \times \frac{364}{365}.$$

When you ask the first person to tell you their birthday, he or she can occupy any of the 365 slots. This gives us $\frac{365}{365}$. The second person has one slot short because the first person has taken it, and so the probability that he or she has a different birthday from the first person is $\frac{364}{365}$. Note that this calculation is independent of how many people you have in the room because you are picking them sequentially.

If you now choose a third person, the probability that they have different birthdays is

$$\mathbb{P}[\text{The first 3 people have different birthdays}] = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365}.$$

This process can be visualized in **Figure 1.17**.

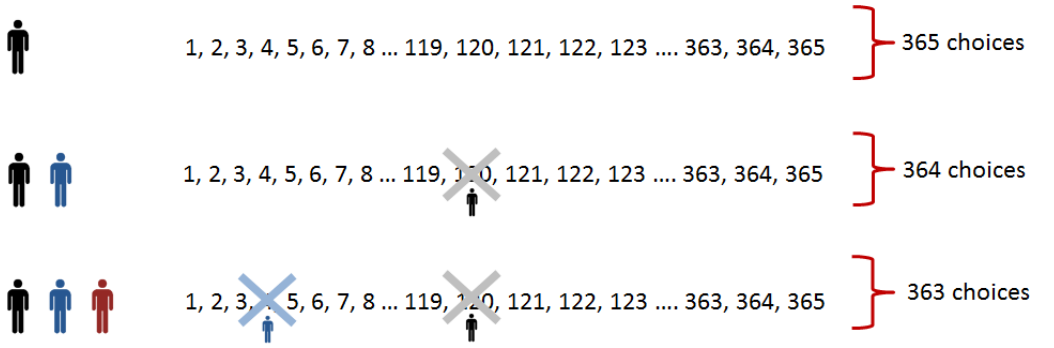


Figure 1.17: The probability for two people to have the same birthday as a function of the number of people in the group. When there is only one person, this person can land on any of the 365 days. When there are two people, the first person has already taken one day (out of 365 days), so the second person can only choose 364 days. When there are three people, the first two people have occupied two days, so there are only 363 days left. If we generalize this process, we see that the number of configurations is $365 \times 364 \times \cdots \times (365 - k + 1)$, where k is the number of people in the room.

So imagine that you keep going down the list to the 50th person. The probability that none of these 50 people will have the same birthday is

$$\begin{aligned}&\mathbb{P}[\text{The first 50 people have different birthdays}]\\ &= \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{316}{365} \approx 0.03.\end{aligned}$$

That means that the probability for 50 people to have different birthdays, the probability is as little as 3%. If you take the complement, you can show that with 97% probability, there is at least one pair of people having the same birthday.

The general equation for this problem is now easy to see:

$$\begin{aligned}\mathbb{P}[\text{The first } k \text{ people have different birthdays}] &= \frac{365 \times 364 \times \cdots \times (365 - k + 1)}{365 \times 365 \times \cdots \times 365}\\ &= \frac{365!}{(365-k)!} \times \frac{1}{365^k}.\end{aligned}$$

The first term in our equation, $\frac{365!}{(365-k)!}$, is called the **permutation** of picking k days from 365 options. We shall discuss this operation shortly.

Why is the probability so high with only 50 people while it seems that we need 366 people to ensure two identical birthdays? The difference is the notion of **probabilistic** and **deterministic**. The 366-people argument is deterministic. If you have 366 people, you are certain that two people will have the same birthday. This has no conflict with the probabilistic argument because the probabilistic argument says that with 50 people, we have a 97% chance of getting two identical birthdays. With a 97% success rate, you still have a 3% chance of failing. It is unlikely to happen, but it can still happen. The more people you put into the room, the stronger guarantee you will have. However, even if you have 364 people and the probability is almost 100%, there is still no guarantee. So there is no conflict between the two arguments since they are answering two different questions.

Now, let's discuss the two combinatorics questions.

1.5.2 Permutation

Permutation concerns the following question:

Consider a set of n distinct balls. Suppose we want to pick k balls from the set without replacement. How many ordered configurations can we obtain?

Note that in the above question, the word "ordered" is crucial. For example, the set $A = \{a, b, c\}$ can lead to 6 different ordered configurations

$$(a, b, c),\ (a, c, b),\ (b, a, c),\ (b, c, a),\ (c, a, b),\ (c, b, a).$$

As a simple illustration of how to compute the permutation, we can consider a set of 5 colored balls as shown in **Figure 1.18**.

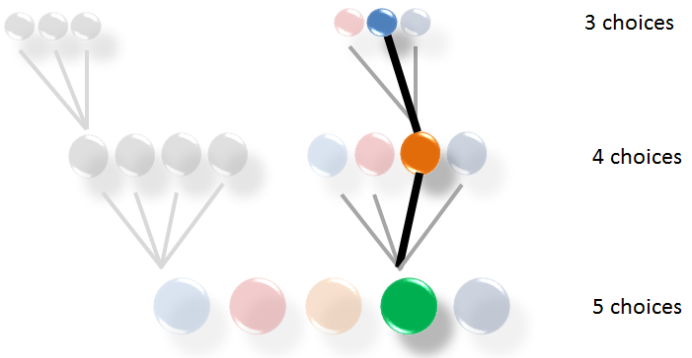


Figure 1.18: Permutation. The number of choices is reduced in every stage. Therefore, the total number is $n \times (n-1) \times \cdots \times (n-k+1)$ if there are k stages.

If you start with the base, which contains five balls, you will have five choices. At one level up, since one ball has already been taken, you have only four choices. You continue the process until you reached the number of balls you want to collect. The number of configurations you have generated is the permutation. Here is the formula:

Theorem 1.6. *The number of* **permutations** *of choosing k out of n is*

$$\frac{n!}{(n-k)!}$$

where $n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$.

Proof. Let's list all possible ways:

Which ball to pick	Number of choices	Why?
The 1st ball	n	No has been picked, so we have n choices
The 2nd ball	$n-1$	The first ball has been picked
The 3rd ball	$n-2$	The first two balls have been picked
$\vdots$	$\vdots$	$\vdots$
The kth ball	$n-k+1$	The first $k-1$ balls have been picked
Total:	$n(n-1)\cdots(n-k+1)$	

The total number of ordered configurations is $n(n-1)\cdots(n-k+1)$. This simplifies to

$$\begin{aligned} &n(n-1)(n-2)\cdots(n-k+1) \\ &= n(n-1)(n-2)\cdots(n-k+1)\cdot\frac{(n-k)(n-k-1)\cdots 3\cdot 2\cdot 1}{(n-k)(n-k-1)\cdots 3\cdot 2\cdot 1} \\ &= \frac{n!}{(n-k)!}. \end{aligned}$$

□

Practice Exercise 1.8. Consider a set of 4 balls $\{1,2,3,4\}$. We want to pick two balls at random without replacement. The ordering matters. How many permutations can we obtain?

Solution. The possible configurations are (1,2), (2,1), (1,3), (3,1), (1,4), (4,1), (2,3), (3,2), (2,4), (4,2), (3,4), (4,3). So totally there are 12 configurations. We can also verify this number by noting that there are 4 balls altogether and so the number of choices for picking the first ball is 4 and the number of choices for picking the second ball is $(4-1)=3$. Thus, the total is $4\cdot 3 = 12$. Referring to the formula, this result coincides with the theorem, which states that the number of permutations is $\frac{4!}{(4-2)!} = \frac{4\cdot 3\cdot 2\cdot 1}{2\cdot 1} = 12$.

1.5.3 Combination

Another operation in combinatorics is combination. Combination concerns the following question:

Consider a set of n distinct balls. Suppose we want to pick k balls from the set without replacement. How many **unordered** *configurations can we obtain?*

Unlike permutation, combination treats a subset of balls with whatever ordering as one single configuration. For example, the subset (a, b, c) is considered the same as (a, c, b) or (b, c, a), etc.

Let's go back to the 5-ball exercise. Suppose you have picked orange, green, and light blue. This is the same combination as if you have picked {green, orange, and light blue}, or {green, light blue, and orange}. **Figure 1.19** lists all the six possible configurations for these three balls. So what is combination? Combination needs to take these repeated cases into account.

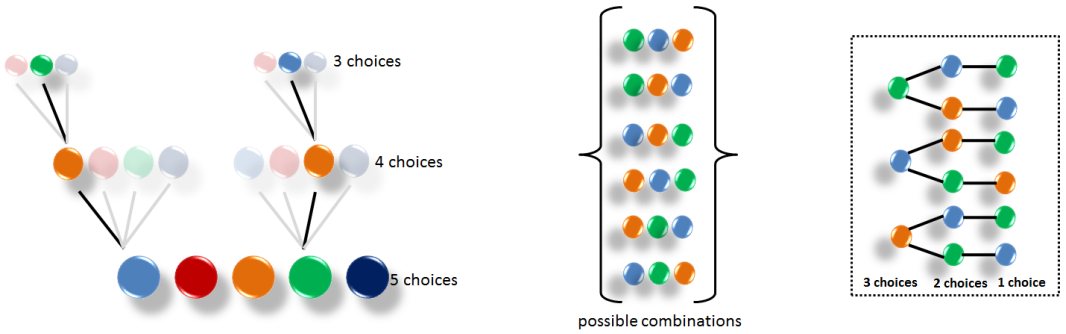


Figure 1.19: Combination. In this problem, we are interested in picking 3 colored balls out of 5. This will give us $5 \times 4 \times 3 = 60$ permutations. However, since we are not interested in the ordering, some of the permutations are repeated. For example, there are 6 combos of (green, light blue, orange), which is computed from $3 \times 2 \times 1$. Dividing 60 permutations by these 6 choices of the orderings will give us 10 distinct combinations of the colors.

Theorem 1.7. *The number of* **combinations** *of choosing k out of n is*

$$\frac{n!}{k!(n-k)!}$$

where $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$.

Proof. We start with the permutation result, which gives us $\frac{n!}{(n-k)!}$ permutations. Note that every permutation has exactly k balls. However, while these k balls can be arranged in any order, in combination, we treat them as one single configuration. Therefore, the task is to count the number of possible orderings for these k balls.

To this end, we note that for a set of k balls, there are in total $k!$ possible ways of ordering them. The number $k!$ comes from the following table.

Which ball to pick	Number of choices
The 1st ball	k
The 2nd ball	$k-1$
$\vdots$	$\vdots$
The kth ball	1
Total:	$k(k-1)\cdots 3\cdot 2\cdot 1$

Therefore, the total number of orderings for a set of k balls is $k!$. Since permutation gives us $\frac{n!}{(n-k)!}$ and every permutation has $k!$ repetitions due to ordering, we divide the number by $k!$. Thus the number of combinations is

$$\frac{n!}{k!(n-k)!}.$$

□

Practice Exercise 1.9. Consider a set of 4 balls $\{1,2,3,4\}$. We want to pick two balls at random without replacement. The ordering does not matter. How many combinations can we obtain?

Solution. The permutation result gives us 12 permutations. However, among all these 12 permutations, there are only 6 distinct pairs of numbers. We can confirm this by noting that since we picked 2 balls, there are exactly 2 possible orderings for these 2 balls. Therefore, we have $\frac{12}{2}=6$ number of combinations. Using the formula of the theorem, we check that the number of combinations is

$$\frac{4!}{2!(4-2)!}=\frac{4\cdot 3\cdot 2\cdot 1}{(2\cdot 1)(2\cdot 1)}=6.$$

Example 1.10. (Ross, 8th edition, Section 1.6) Consider the equation

$$x_1+x_2+\cdots+x_K=N,$$

where $\{x_k\}$ are positive integers. How many combinations of solutions of this equation are there?

Solution. We can determine the number of combinations by considering the figure below. The integer N can be modeled as N balls in an urn. The number of variables K is equivalent to the number of colors of these balls. Since all variables are positive, the problem can be translated to partitioning the N balls into K buckets. This, in turn, is the same as inserting $K-1$ dividers among $N-1$ holes. Therefore, the number of combinations is

$$\binom{N-1}{K-1}=\frac{(N-1)!}{(K-1)!(N-K)!}.$$

For example, if $N=16$ and $K=4$, then the number of solutions is

$$\binom{16-1}{4-1}=\frac{15!}{3!12!}=455.$$

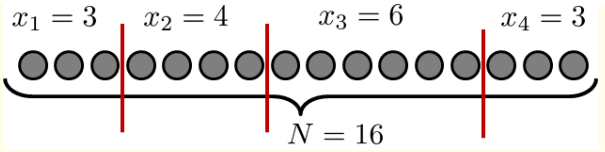


Figure 1.20: One possible solution for $N = 16$ and $K = 4$. In general, the problem is equivalent to inserting $K - 1$ dividers among $N - 1$ balls.

Closing remark. Permutations and combinations are two ways to enumerate all the possible cases. While the conclusions are probabilistic, as the birthday paradox shows, permutation and combination are deterministic. We do not need to worry about the distribution of the samples, and we are not taking averages of anything. Thus, modern data analysis seldom uses the concepts of permutation and combination. Accordingly, combinatorics does not play a large role in this book.

Does it mean that combinatorics is not useful? Not quite, because it still provides us with powerful tools for theoretical analysis. For example, in binomial random variables, we need the concept of combination to calculate the repeated cases. The Poisson random variable can be regarded as a limiting case of the binomial random variable, and so combination is also used. Therefore, while we do not use the concepts of permutation per se, we use them to define random variables.

1.6 Summary

In this chapter, we have reviewed several background mathematical concepts that will become useful later in the book. You will find that these concepts are important for understanding the rest of this book. When studying these materials, we recommend not just remembering the "recipes" of the steps but focusing on the **motivations** and **intuitions** behind the techniques.

We would like to highlight the significance of the birthday paradox. Many of us come from an engineering background in which we were told to ensure reliability and guarantee success. We want to ensure that the product we deliver to our customers can survive even in the worst-case scenario. We tend to apply deterministic arguments such as requiring 366 people to ensure complete coverage of the 365 days. In modern data analysis, the worst-case scenario may not always be relevant because of the complexity of the problem and the cost of such a warranty. The probabilistic argument, or the average argument, is more reasonable and cost-effective, as you can see from our analysis of the birthday problem. The heart of the problem is the trade-off between how much confidence you need versus how much effort you need to expend. Suppose an event is unlikely to happen, but if it happens, it will be a disaster. In that case, you might prefer to be very conservative to ensure that such a disaster event has a low chance of happening. Industries related to risk management such as insurance and investment banking are all operating under this principle.

1.7 Reference

Introductory materials

1-1 Erwin Kreyszig, *Advanced Engineering Mathematics*, Wiley, 10th Edition, 2011.

1-2 Henry Stark and John W. Woods, *Probability and Random Processes with Applications to Signal Processing*, Prentice Hall, 3rd Edition, 2002. Appendix.

1-3 Michael J. Evans and Jeffrey S. Rosenthal, *Probability and Statistics: The Science of Uncertainty*, W. H. Freeman, 2nd Edition, 2009. Appendix.

1-4 James Stewart, *Single Variable Calculus, Early Transcendentals*, Thomson Brooks/-Cole, 6th Edition, 2008. Chapter 5.

Combinatorics

1-5 Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd Edition, 2008. Section 1.6.

1-6 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Section 2.6.

1-7 Athanasios Papoulis and S. Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th Edition, 2001. Chapter 3.

Analysis

In some sections of this chapter, we use results from calculus and infinite series. Many formal proofs can be found in the standard undergraduate real analysis textbooks.

1-8 Tom M. Apostol, *Mathematical Analysis*, Pearson, 1974.

1-9 Walter Rudin, *Principles of Mathematical Analysis*, McGraw Hill, 1976.

1.8 Problems

Exercise 1. (Video Solution)

(a) Show that

$$\sum_{k=0}^{n} r^k = \frac{1 - r^{n+1}}{1 - r}.$$

for any $0 < r < 1$. Evaluate $\sum_{k=0}^{\infty} r^k$.

(b) Using the result of (a), evaluate

$$1 + 2r + 3r^2 + \cdots.$$

(c) Evaluate the sums

$$\sum_{k=0}^{\infty} k \left(\frac{1}{3}\right)^{k+1}, \text{ and } \sum_{k=2}^{\infty} k \left(\frac{1}{4}\right)^{k-1}.$$

Exercise 2. (Video Solution)
Recall that

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}.$$

Evaluate

$$\sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!}, \text{ and } \sum_{k=0}^{\infty} k^2 \frac{\lambda^k e^{-\lambda}}{k!}.$$

Exercise 3. (Video Solution)
Evaluate the integrals

(a)

$$\int_a^b \frac{1}{b-a} \left(x - \frac{a+b}{2}\right)^2 \, dx.$$

(b)

$$\int_0^{\infty} \lambda x e^{-\lambda x} \, dx.$$

(c)

$$\int_{-\infty}^{\infty} \frac{\lambda x}{2} e^{-\lambda |x|} \, dx.$$

Exercise 4.

(a) Compute the result of the following matrix vector multiplication using Numpy. Submit your result and codes.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

(b) Plot a sine function on the interval $[-\pi, \pi]$ with 1000 data points.

(c) Generate 10,000 uniformly distributed random numbers on interval [0, 1).

Use `matplotlib.pyplot.hist` to generate a histogram of all the random numbers.

Exercise 5.
Calculate

$$\sum_{k=0}^{\infty} k \left(\frac{2}{3}\right)^{k+1}.$$

Exercise 6.
Let

$$\boldsymbol{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix}.$$

(a) Find $\boldsymbol{\Sigma}^{-1}$, the inverse of $\boldsymbol{\Sigma}$.

(b) Find $|\boldsymbol{\Sigma}|$, the determinant of $\boldsymbol{\Sigma}$.

(c) Simplify the two-dimensional function

$$f(\boldsymbol{x}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

(d) Use `matplotlib.pyplot.contour`, plot the function $f(\boldsymbol{x})$ for the range $[-3,3] \times [-3,3]$.

Exercise 7.
Out of seven electrical engineering (EE) students and five mechanical engineering (ME) students, a committee consisting of three EEs and two MEs is to be formed. In how many ways can this be done if

(a) any of the EEs and any of the MEs can be included?

(b) one particular EE must be on the committee?

(c) two particular MEs cannot be on the committee?

Exercise 8.
Five blue balls, three red balls, and three white balls are placed in an urn. Three balls are drawn at random without regard to the order in which they are drawn. Using the counting approach to probability, find the probability that

(a) one blue ball, one red ball, and one white ball are drawn.

(b) all three balls drawn are red.

(c) exactly two of the balls drawn are blue.

Exercise 9.
A collection of 26 English letters, a-z, is mixed in a jar. Two letters are drawn at random, one after the other.

(a) What is the probability of drawing a vowel (a,e,i,o,u) and a consonant in either order?

(b) Write a MATLAB / Python program to verify your answer in part (a). Randomly draw two letters without replacement and check whether one is a vowel and the other is a consonant. Compute the probability by repeating the experiment 10000 times.

Exercise 10.
There are 50 students in a classroom.

(a) What is the probability that there is at least one pair of students having the same birthday? Show your steps.

(b) Write a MATLAB / Python program to simulate the event and verify your answer in (a). Hint: You probably need to repeat the simulation many times to obtain a probability. Submit your code and result.

You may assume that a year only has 365 days. You may also assume that all days have an equal likelihood of being taken.

Chapter 2

Probability

Data and probability are inseparable. Data is the **computational** side of the story, whereas probability is the **theoretical** side of the story. Any data science practice must be built on the foundation of probability, and probability needs to address practical problems. However, what exactly is "probability"? Mathematicians have been debating this for centuries. The **frequentists** argue that probability is the relative frequency of an outcome. For example, flipping a fair coin has a 1/2 probability of getting a head because if you flip the coin infinitely many times, you will have half of the time getting a head. The **Bayesians** argue that probability is a subjective belief. For example, the probability of getting an A in a class is subjective because no one would want to take a class infinitely many times to obtain the relative frequency. Both the frequentists and Bayesians have valid points. However, the differentiation is often non-essential because the context of your problem will force you to align with one or the other. For example, when you have a shortage of data, then the subjectivity of the Bayesians allows you to use prior knowledge, whereas the frequentists tell us how to compute the confidence interval of an estimate.

No matter whether you prefer the frequentist's view or the Bayesian's view, there is something more fundamental thanks to **Andrey Kolmogorov** (1903-1987). The development of this fundamental definition will take some effort on our part, but if we distill the essence, we can summarize it as follows:

Probability is a measure of the size of a set.

This sentence is not a formal definition; instead, it summarizes what we believe to be the essence of probability. We need to clarify some puzzles later in this chapter, but if you can understand what this sentence means, you are halfway done with this book. To spell out the details, we will describe an elementary problem that everyone knows how to solve. As we discuss this problem, we will highlight a few key concepts that will give you some intuitive insights into our definition of probability, after which we will explain the sequence of topics to be covered in this chapter.

Prelude: Probability of throwing a die

Suppose that you have a fair die. It has 6 faces: {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}. What is the probability that you get a number that is "less than 5" and is "an even number"? This is a straightfor-

ward problem. You probably have already found the answer, which is $\frac{2}{6}$ because "less than 5" and "an even number" means $\{⚀, ⚁\}$. However, let's go through the thinking process slowly by explicitly writing down the steps.

First of all, how do we know that the denominator in $\frac{2}{6}$ is 6? Well, because there are six faces. These six faces form a set called the **sample space**. A sample space is the set containing all possible outcomes, which in our case is $\Omega = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\}$. The denominator 6 is the size of the sample space.

How do we know that the numerator is 2? Again, implicitly in our minds, we have constructed two **events**: E_1 = "less than 5" = $\{⚀, ⚁, ⚂, ⚃\}$, and E_2 = "an even number" = $\{⚁, ⚃, ⚅\}$. Then we take the intersection between these two events to conclude the event $E = \{⚀, ⚁\}$. The numerical value "2" is the size of this event E.

So, when we say that "the probability is $\frac{2}{6}$," we are saying that the size of the event E relative to the sample space Ω is the ratio $\frac{2}{6}$. This process involves **measuring** the size of E and Ω. In this particular example, the measure we use is a "counter" that counts the number of elements.

This example shows us all the necessary components of probability: (i) There is a **sample space**, which is the set that contains all the possible outcomes. (ii) There is an **event**, which is a subset inside the sample space. (iii) Two events E_1 and E_2 can be **combined** to construct another event E that is still a subset inside the sample space. (iv) Probability is a number assigned by certain **rules** such that it describes the **relative size** of the event E compared with the sample space Ω. So, when we say that **probability is a measure of the size of a set**, we create a mapping that takes in a set and outputs the size of that set.

Organization of this chapter

As you can see from this example, since probability is a measure of the size of a set, we need to understand the operations of sets to understand probability. Accordingly, in Section 2.1 we first define sets and discuss their operations. After learning these basic concepts, we move on to define the sample space and event space in Section 2.2. There, we discuss sample spaces that are not necessarily countable and how probabilities are assigned to events. Of course, assigning a probability value to an event cannot be arbitrary; otherwise, the probabilities may be inconsistent. Consequently, in Section 2.3 we introduce the probability axioms and formalize the notion of measure. Section 2.4 consists of a trio of topics that concern the relationship between events using conditioning. We discuss conditional probability in Section 2.4.1, independence in Section 2.4.2, and Bayes' theorem in Section 2.4.3.

2.1 Set Theory

2.1.1 Why study set theory?

In mathematics, we are often interested in describing a collection of numbers, for example, a positive interval $[a, b]$ on the real line or the ordered pairs of numbers that define a circle on a graph with two axes. These collections of numbers can be abstractly defined as **sets**. In a nutshell, a set is simply a collection of things. These things can be numbers, but they can also be alphabets, objects, or anything. Set theory is a mathematical tool that defines operations on sets. It provides the basic arithmetic for us to combine, separate, and decompose sets.

Why do we start the chapter by describing set theory? Because **probability is a measure of the size of a set**. Yes, probability is not just a number telling us the relative frequency of events; it is an operator that takes a set and tells us how large the set is. Using the example we showed in the prelude, the event "even number" of a die is a set containing numbers $\{⚁, ⚃, ⚅\}$. When we apply probability to this set, we obtain the number $\frac{3}{6}$, as shown in **Figure 2.1**. Thus sets are the foundation of the study of probability.

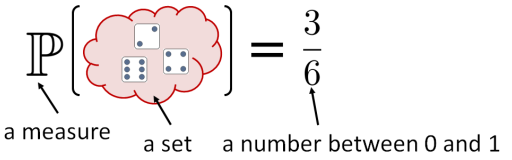


Figure 2.1: Probability is a measure of the size of a set. Whenever we talk about probability, it has to be the probability of a **set**.

2.1.2 Basic concepts of a set

Definition 2.1 (Set). *A* **set** *is a collection of elements. We denote*

$$A = \{\xi_1, \xi_2, \ldots, \xi_n\} \tag{2.1}$$

as a set, where ξ_i is the ith element in the set.

In this definition, A is called a set. It is nothing but a collection of elements $\xi_1, \ldots, \xi_n$. What are these ξ_i's? They can be anything. Let's see a few examples below.

Example 2.1(a). $A = \{\text{apple}, \text{orange}, \text{pear}\}$ is a finite set.
Example 2.1(b). $A = \{1, 2, 3, 4, 5, 6\}$ is a finite set.
Example 2.1(c). $A = \{2, 4, 6, 8, \ldots\}$ is a countable but infinite set.
Example 2.1(d). $A = \{x \mid 0 < x < 1\}$ is a uncountable set.

To say that an element ξ is drawn from A, we write $\xi \in A$. For example, the number 1 is an element in the set $\{1, 2, 3\}$. We write $1 \in \{1, 2, 3\}$. There are a few common sets that we will encounter. For example,

Example 2.2(a). $\mathbb{R}$ is the set of all real numbers including $\pm\infty$.
Example 2.2(b). $\mathbb{R}^2$ is the set of ordered pairs of real numbers.
Example 2.2(c). $[a, b] = \{x \,|\, a \leq x \leq b\}$ is a closed interval on $\mathbb{R}$.
Example 2.2(d). $(a, b) = \{x \,|\, a < x < b\}$ is an open interval on $\mathbb{R}$.
Example 2.2(e). $(a, b] = \{x \,|\, a < x \leq b\}$ is a semi-closed interval on $\mathbb{R}$.

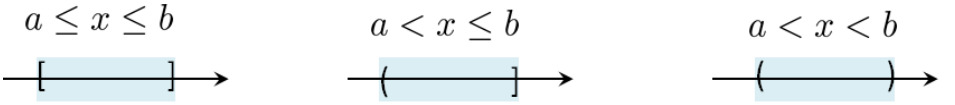


Figure 2.2: From left to right: a closed interval, a semi-closed (or semi-open) interval, and an open interval.

Sets are not limited to numbers. A set can be used to describe a collection of **functions**.

Example 2.3. $A = \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = ax+b,\ a, b \in \mathbb{R}\}$. This is the set of all straight lines in 2D. The notation $f : \mathbb{R} \to \mathbb{R}$ means that the function f takes an argument from $\mathbb{R}$ and sends it to another real number in $\mathbb{R}$. The definition $f(x) = ax + b$ says that f is taking the specific form of $ax + b$. Since the constants a and b can be any real number, the equation $f(x) = ax + b$ enumerates all possible straight lines in 2D. See **Figure 2.3**(a).

Example 2.4. $A = \{f : \mathbb{R} \to [-1, 1] \mid f(t) = \cos(\omega_0 t + \theta),\ \theta \in [0, 2\pi]\}$. This is the set of all cosine functions of a fixed carrier frequency ω_0. The phase θ, however, is changing. Therefore, the equation $f(t) = \cos(\omega_0 t + \theta)$ says that the set A is the collection of all possible cosines with different phases. See **Figure 2.3**(b).

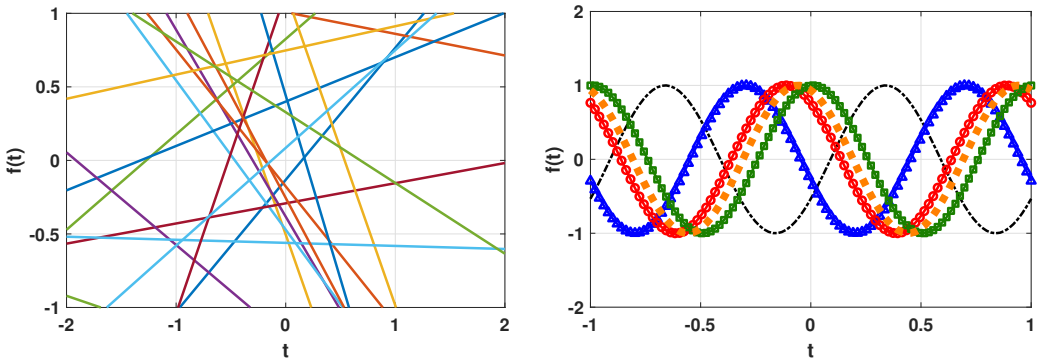


Figure 2.3: (a) The set of straight lines $A = \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = ax + b,\ a, b \in \mathbb{R}\}$. (b) The set of phase-shifted cosines $A = \{f : \mathbb{R} \to [-1, 1] \mid f(t) = \cos(\omega_0 t + \theta),\ \theta \in [0, 2\pi]\}$.

A set can also be used to describe a collection of sets. Let A and B be two sets. Then $\mathcal{C} = \{A, B\}$ is a set of sets.

Example 2.5. Let $A = \{1, 2\}$ and $B = \{\text{apple}, \text{orange}\}$. Then

$$\mathcal{C} = \{A, B\} = \{\{1, 2\}, \{\text{apple}, \text{orange}\}\}$$

is a collection of sets. Note that here we are not saying $\mathcal{C}$ is the union of two sets. We are only saying that $\mathcal{C}$ is a collection of two sets. See the next example.

Example 2.6. Let $A = \{1, 2\}$ and $B = \{3\}$, then $\mathcal{C} = \{A, B\}$ means that

$$\mathcal{C} = \{\{1, 2\}, \{3\}\}.$$

Therefore $\mathcal{C}$ contains only two elements. One is the set $\{1, 2\}$ and the other is the set $\{3\}$. Note that $\{\{1, 2\}, \{3\}\} \neq \{1, 2, 3\}$. The former is a set of two sets. The latter is a set of three elements.

2.1.3 Subsets

Given a set, we often want to specify a portion of the set, which is called a **subset**.

Definition 2.2 (**Subset**). *B is a **subset** of A if for any $\xi \in B$, ξ is also in A. We write*

$$B \subseteq A \tag{2.2}$$

to denote that B is a subset of A.

B is called a **proper subset** of A if B is a subset of A and $B \neq A$. We denote a proper subset as $B \subset A$. Two sets A and B are equal if and only if $A \subseteq B$ and $B \subseteq A$.

Example 2.7.

- If $A = \{1, 2, 3, 4, 5, 6\}$, then $B = \{1, 3, 5\}$ is a proper subset of A.
- If $A = \{1, 2\}$, then $B = \{1, 2\}$ is an improper subset of A.
- If $A = \{t \mid t \geq 0\}$, then $B = \{t \mid t > 0\}$ is a proper subset of A.

Practice Exercise 2.1. Let $A = \{1, 2, 3\}$. List all the subsets of A.
Solution. The subsets of A are:

$$\mathcal{A} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

Practice Exercise 2.2. Prove that two sets A and B are equal if and only if $A \subseteq B$ and $B \subseteq A$.

Solution. Suppose $A \subseteq B$ and $B \subseteq A$. Assume by contradiction that $A \neq B$. Then necessarily there must exist an x such that $x \in A$ but $x \notin B$ (or vice versa). But $A \subseteq B$ means that $x \in A$ will necessarily be in B. So it is impossible to have $x \notin B$. Conversely, suppose that $A = B$. Then any $x \in A$ will necessarily be in B. Therefore, we have $A \subseteq B$. Similarly, if $A = B$ then any $x \in B$ will be in A, and so $B \subseteq A$.

2.1.4 Empty set and universal set

Definition 2.3 (**Empty Set**). *A set is* **empty** *if it contains no element. We denote an empty set as*

$$A = \emptyset. \tag{2.3}$$

A set containing an element 0 is not an empty set. It is a set of one element, $\{0\}$. The number of elements of the empty set is 0. The empty set is a subset of any set, i.e., $\emptyset \subseteq A$ for any A. We use $\subseteq$ because A could also be an empty set.

Example 2.8(a). The set $A = \{x \mid \sin x > 1\}$ is empty because no $x \in \mathbb{R}$ can make $\sin x > 1$.

Example 2.8(b). The set $A = \{x \mid x > 5 \text{ and } x < 1\}$ is empty because the two conditions $x > 5$ and $x < 1$ are contradictory.

Definition 2.4 (**Universal Set**). *The* **universal set** *is the set containing all elements under consideration. We denote a universal set as*

$$A = \Omega. \tag{2.4}$$

The universal set Ω contains itself, i.e., $\Omega \subseteq \Omega$. The universal set is a relative concept. Usually, we first define a universal set Ω before referring to subsets of Ω. For example, we can define $\Omega = \mathbb{R}$ and refer to intervals in $\mathbb{R}$. We can also define $\Omega = [0, 1]$ and refer to subintervals inside $[0, 1]$.

2.1.5 Union

We now discuss basic set operations. By operations, we mean functions of two or more sets whose output value is a set. We use these operations to combine and separate sets. Let us first consdier the union of two sets. See **Figure 2.4** for a graphical depiction.

Definition 2.5 (**Finite Union**). *The* **union** *of two sets A and B contains all elements in A* **or** *in B. That is,*

$$A \cup B = \{\xi \mid \xi \in A \textbf{ or } \xi \in B\}. \tag{2.5}$$

As the definition suggests, the union of two sets connects the sets using the logical operator "**or**". Therefore, the union of two sets is always larger than or equal to the individual sets.

Example 2.9(a). If $A = \{1, 2\}$, $B = \{1, 5\}$, then $A \cup B = \{1, 2, 5\}$. The overlapping element 1 is absorbed. Also, note that $A \cup B \neq \{\{1, 2\}, \{1, 5\}\}$. The latter is a set of sets.

Example 2.9(b). If $A = (3, 4]$, $B = (3.5, \infty)$, then $A \cup B = (3, \infty)$.

Example 2.9(c). If $A = \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = ax\}$ and $B = \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = b\}$, then $A \cup B =$ a set of sloped lines with a slope a plus a set of constant lines with

height b. Note that $A \cup B \neq \{f : \mathbb{R} \to \mathbb{R} \,|\, f(x) = ax + b\}$ because the latter is a set of sloped lines with arbitrary y-intercept.

Example 2.9(d). If $A = \{1, 2\}$ and $B = \emptyset$, then $A \cup B = \{1, 2\}$.

Example. If $A = \{1, 2\}$ and $B = \Omega$, then $A \cup B = \Omega$.

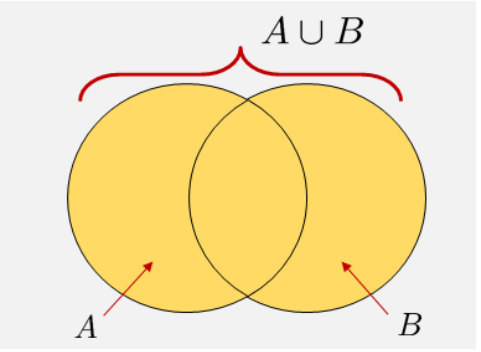


Figure 2.4: The union of two sets contains elements that are either in A or B or both.

The previous example can be generalized in the following exercise. What it says is that if A is a subset of another set B, then the union of A and B is just B. Intuitively, this should be straightforward because whatever you have in A is already in B, so the union will just be B. Below is a formal proof that illustrates how to state the arguments clearly. You may like to draw a picture to convince yourself that the proof is correct.

Practice Exercise 2.3: Prove that if $A \subseteq B$, then $A \cup B = B$.

Solution: We will show that $A \cup B \subseteq B$ and $B \subseteq A \cup B$. Let $\xi \in A \cup B$. Then ξ must be inside either A or B (or both). In any case, since we know that $A \subseteq B$, it holds that if $\xi \in A$ then ξ must also be in B. Therefore, for any $\xi \in A \cup B$ we have $\xi \in B$. This shows $A \cup B \subseteq B$. Conversely, if $\xi \in B$, then ξ must be inside $A \cup B$ because $A \cup B$ is a larger set than B. So if $\xi \in B$ then $\xi \in A \cup B$ and hence $B \subseteq A \cup B$. Since $A \cup B$ is a subset of B or equal to B, and B is a subset of $A \cup B$ or equal to $A \cup B$, it follows that $A \cup B = B$.

What should we do if we want to take the union of an infinite number of sets? First, we need to define the concept of an **infinite union**.

Definition 2.6 (**Infinite Union**). *For an infinite sequence of sets $A_1, A_2, \ldots$, the* **infinite union** *is defined as*

$$\bigcup_{n=1}^{\infty} A_n = \{\xi \mid \xi \in A_n \text{ for } \textbf{at least one } n \text{ that is finite.}\}. \tag{2.6}$$

An infinite union is a natural extension of a finite union. It is not difficult to see that

$$\xi \in A \text{ or } \xi \in B \iff \xi \text{ is in } \textbf{at least one of } A \text{ and } B.$$

Similarly, an infinite union means that

$$\xi \in A_1 \text{ or } \xi \in A_2 \text{ or } \xi \in A_3 \ldots \iff \xi \text{ is in at least one of } A_1,\ A_2,\ A_3,\ \ldots.$$

The finite n requirement says that we only evaluate the sets for a finite number of n's. This n can be arbitrarily large, but it is finite. Why are we able to do this? Because the concept of an infinite union is to determine A_∞, which is the limit of a sequence. Like any sequence of real numbers, the limit of a sequence of sets has to be defined by evaluating the instances of all possible finite cases.

Consider a sequence of sets $A_n = \left[-1, 1 - \frac{1}{n}\right]$, for $n = 1, 2, \ldots$. For example, $A_1 = [-1, 0]$, $A_2 = \left[-1, \frac{1}{2}\right]$, $A_3 = \left[-1, \frac{2}{3}\right]$, $A_4 = \left[-1, \frac{3}{4}\right]$, etc.

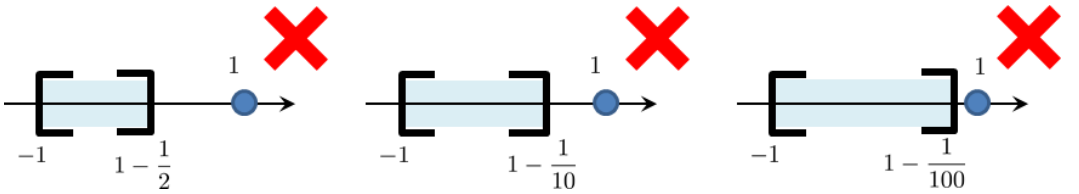


Figure 2.5: The infinite union of $\bigcup_{n=1}^{\infty}\left[-1, 1 - \frac{1}{n}\right]$. No matter how large n gets, the point 1 is never included. So the infinite union is $[-1, 1)$

To take the infinite union, we know that the set $[-1, 1)$ is always included, because the right-hand limit $1 - \frac{1}{n}$ approaches 1 as n approaches ∞. So the only question concerns the number 1. Should 1 be included? According to the definition above, we ask: Is 1 an element of **at least one** of the sets $A_1, A_2, \ldots, A_n$? Clearly it is not: $1 \notin A_1$, $1 \notin A_2$, ... In fact, $1 \notin A_n$ for any finite n. Therefore 1 is not an element of the infinite union, and we conclude that

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty}\left[-1, 1 - \frac{1}{n}\right] = [-1, 1).$$

Practice Exercise 2.4. Find the infinite union of the sequences where (a) $A_n = \left[-1, 1 - \frac{1}{n}\right)$, (b) $A_n = \left(-1, 1 - \frac{1}{n}\right]$.
Solution. (a) $\bigcup_{n=1}^{\infty} A_n = [-1, 1)$. (b) $\bigcup_{n=1}^{\infty} A_n = (-1, 1)$.

2.1.6 Intersection

The union of two sets is based on the logical operator **or**. If we use the logical operator **and**, then the result is the **intersection** of two sets.

Definition 2.7 (**Finite Intersection**). *The* **intersection** *of two sets A and B contains all elements in A* **and** *in B. That is,*

$$A \cap B = \{\xi \mid \xi \in A \text{ and } \xi \in B\}. \tag{2.7}$$

Figure 2.6 portrays intersection graphically. Intersection finds the common elements of the two sets. It is not difficult to show that $A \cap B \subseteq A$ and $A \cap B \subseteq B$.

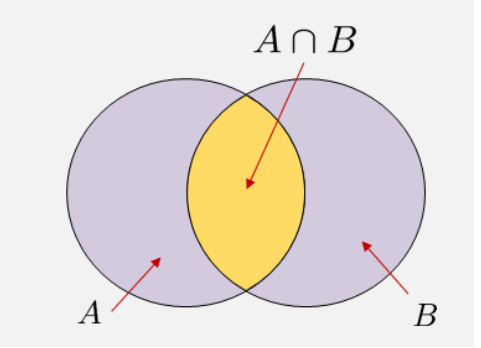


Figure 2.6: The intersection of two sets contains elements in both A and B.

Example 2.10(a). If $A = \{1, 2, 3, 4\}$, $B = \{1, 5, 6\}$, then $A \cap B = \{1\}$.

Example 2.10(b). If $A = \{1, 2\}$, $B = \{5, 6\}$, then $A \cap B = \emptyset$.

Example 2.10(c). If $A = (3, 4]$, $B = [3.5, \infty)$, then $A \cap B = [3.5, 4]$.

Example 2.10(d). If $A = (3, 4]$, $B = \emptyset$, then $A \cap B = \emptyset$.

Example 2.10(e). If $A = (3, 4]$, $B = \Omega$, then $A \cap B = (3, 4]$.

Example 2.11. If $A = \{f : \mathbb{R} \to \mathbb{R} \,|\, f(x) = ax\}$ and $B = \{f : \mathbb{R} \to \mathbb{R} \,|\, f(x) = b\}$, then $A \cap B =$ the intersection of a set of sloped lines with a slope a and a set of constant lines with height b. The only line that can satisfy both sets is the line $f(x) = 0$. Therefore, $A \cap B = \{f \,|\, f(x) = 0\}$.

Example 2.12. If $A = \{\{1\}, \{2\}\}$ and $B = \{\{2, 3\}, \{4\}\}$, then $A \cap B = \emptyset$. This is because A is a set containing two sets, and B is a set containing two sets. The two sets $\{2\}$ and $\{2, 3\}$ are not the same. Thus, A and B have no elements in common, and so $A \cap B = \emptyset$.

Similarly to the infinite union, we can define the concept of **infinite intersection**.

Definition 2.8 (**Infinite Intersection**). *For an infinite sequence of sets $A_1, A_2, \ldots$, the* **infinite intersection** *is defined as*

$$\bigcap_{n=1}^{\infty} A_n = \{\xi \mid \xi \in A_n \text{ for every } \textit{finite } n.\} \tag{2.8}$$

To understand this definition, we note that

$$\xi \in A \text{ and } \xi \in B \iff \xi \text{ is in every one of } A \text{ and } B.$$

As a result, it follows that

$$\xi \in A_1 \text{ and } \xi \in A_2 \text{ and } \xi \in A_3 \ldots \iff \xi \text{ is in every one of } A_1,\ A_2,\ A_3,\ \ldots.$$

Since the infinite intersection requires that ξ is in every one of A_1, A_2, ..., A_n, if there is a set A_i that does not contain ξ, the infinite intersection is an empty set.

Consider the problem of finding the infinite intersection of $\bigcap_{n=1}^{\infty} A_n$, where

$$A_n = \left[0, 1 + \frac{1}{n}\right).$$

We note that the sequence of sets is $[0, 2]$, $[0, 1.5]$, $[0, 1.33]$, ... As $n \to \infty$, we note that the limit is either $[0, 1)$ or $[0, 1]$. Should the right-hand limit 1 be included in the infinite intersection? According to the definition above, we know that $1 \in A_1$, $1 \in A_2$, ..., $1 \in A_n$ for any finite n. Therefore, 1 is included and so

$$\bigcap_{n=1}^{\infty} A_n = \bigcap_{n=1}^{\infty} \left[0, 1 + \frac{1}{n}\right) = [0, 1].$$

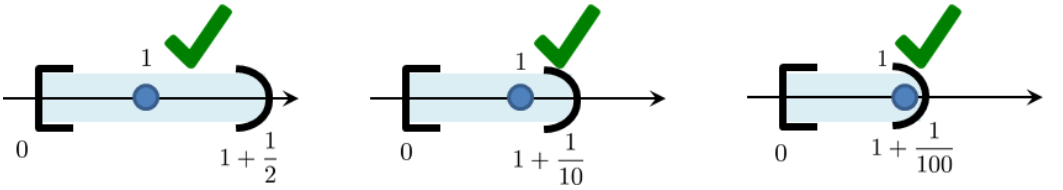


Figure 2.7: The infinite intersection of $\bigcap_{n=1}^{\infty} \left[0, 1 + \frac{1}{n}\right)$. No matter how large n gets, the point 1 is never included. So the infinite intersection is $[0, 1]$

Practice Exercise 2.5. Find the infinite intersection of the sequences where (a) $A_n = \left[0, 1 + \frac{1}{n}\right]$, (b) $A_n = \left(0, 1 + \frac{1}{n}\right)$, (c) $A_n = \left[0, 1 - \frac{1}{n}\right)$, (d) $A_n = \left[0, 1 - \frac{1}{n}\right]$.

Solution.

(a) $\bigcap_{n=1}^{\infty} A_n = [0, 1]$.

(b) $\bigcap_{n=1}^{\infty} A_n = (-1, 1]$.

(c) $\bigcap_{n=1}^{\infty} A_n = [0, 0) = \emptyset$.

(d) $\bigcap_{n=1}^{\infty} A_n = [0, 0] = \{0\}$.

2.1.7 Complement and difference

Besides union and intersection, there is a third basic operation on sets known as the **complement**.

Definition 2.9 (**Complement**). *The* **complement** *of a set A is the set containing all elements that are in Ω but not in A. That is,*

$$A^c = \{\xi \mid \xi \in \Omega \text{ and } \xi \notin A\}. \tag{2.9}$$

Figure 2.8 graphically portrays the idea of a complement. The complement is a set that contains everything in the universal set that is not in A. Thus the complement of a set is always relative to a specified universal set.

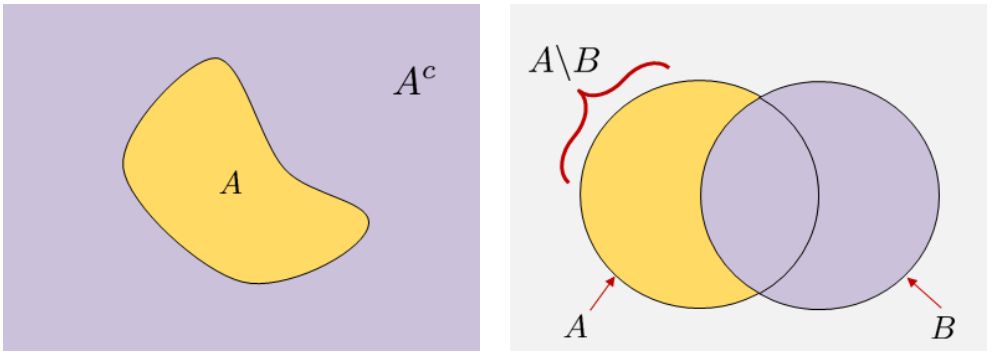


Figure 2.8: [Left] The complement of a set A contains all elements that are not in A. [Right] The difference $A\backslash B$ contains elements that are in A but not in B.

Example 2.13(a). Let $A = \{1,2,3\}$ and $\Omega = \{1,2,3,4,5,6\}$. Then $A^c = \{4,5,6\}$.

Example 2.13(b). Let $A = \{\text{even integers}\}$ and $\Omega = \{\text{integers}\}$. Then $A^c = \{\text{odd integers}\}$.

Example 2.13(c). Let $A = \{\text{integers}\}$ and $\Omega = \mathbb{R}$. Then $A^c = \{\text{any real number that is not an integer}\}$.

Example 2.13(d). Let $A = [0, 5)$ and $\Omega = \mathbb{R}$. Then $A^c = (-\infty, 0) \cup [5, \infty)$.

Example 2.13(e). Let $A = \mathbb{R}$ and $\Omega = \mathbb{R}$. Then $A^c = \emptyset$.

The concept of the complement will help us understand the concept of **difference**.

Definition 2.10 (**Difference**). *The* **difference** $A\backslash B$ *is the set containing all elements in* A *but not in* B.

$$A\backslash B = \{\xi \mid \xi \in A \text{ and } \xi \notin B\}. \tag{2.10}$$

Figure 2.8 portrays the concept of difference graphically. Note that $A\backslash B \neq B\backslash A$. The former removes the elements in B whereas the latter removes the elements in A.

Example 2.14(a). Let $A = \{1, 3, 5, 6\}$ and $B = \{2, 3, 4\}$. Then $A\backslash B = \{1, 5, 6\}$ and $B\backslash A = \{2, 4\}$.

Example 2.14(b). Let $A = [0, 1]$, $B = [2, 3]$, then $A\backslash B = [0, 1]$, and $B\backslash A = [2, 3]$. This example shows that if the two sets do not overlap, there is nothing to subtract.

Example 2.14(c). Let $A = [0, 1]$, $B = \mathbb{R}$, then $A\backslash B = \emptyset$, and $B\backslash A = (-\infty, 0)\cup(1, \infty)$. This example shows that if one of the sets is the universal set, then the difference will either return the empty set or the complement.

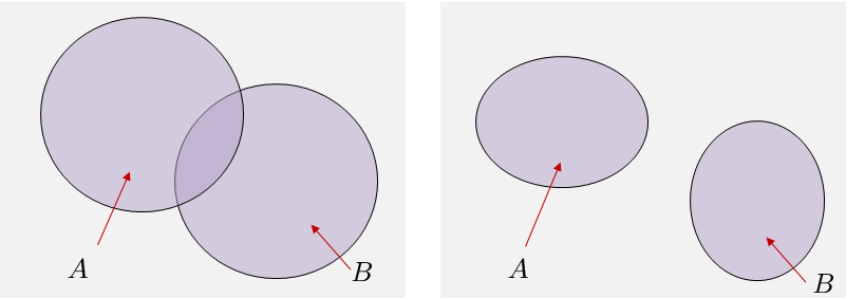


Figure 2.9: [Left] A and B are overlapping. [Right] A and B are disjoint.

Practice Exercise 2.6. Show that for any two sets A and B, the differences $A\backslash B$ and $B\backslash A$ never overlap, i.e., $(A\backslash B) \cap (B\backslash A) = \emptyset$.

Solution. Suppose, by contradiction, that the intersection is not empty so that there exists an $\xi \in (A\backslash B) \cap (B\backslash A)$. Then, by the definition of intersection, ξ is an element of $(A\backslash B)$ **and** $(B\backslash A)$. But if ξ is an element of $(A\backslash B)$, it cannot be an element of B. This implies that ξ cannot be an element of $(B\backslash A)$ since it is a subset of B. This is a contradiction because we just assumed that the ξ can live in both $(A\backslash B)$ and $(B\backslash A)$.

Difference can be defined in terms of intersection and complement:

Theorem 2.1. *Let A and B be two sets. Then*

$$A\backslash B = A \cap B^c \tag{2.11}$$

Proof. Let $x \in A\backslash B$. Then $x \in A$ and $x \notin B$. Since $x \notin B$, we have $x \in B^c$. Therefore, $x \in A$ and $x \in B^c$. By the definition of intersection, we have $x \in A \cap B^c$. This shows that $A\backslash B \subseteq A \cap B^c$. Conversely, let $x \in A \cap B^c$. Then, $x \in A$ and $x \in B^c$, which implies that $x \in A$ and $x \notin B$. By the definition of $A\backslash B$, we have that $x \in A\backslash B$. This shows that $A \cap B^c \subseteq A\backslash B$.

□

2.1.8 Disjoint and partition

It is important to be able to quantify situations in which two sets are not overlapping. In this situation, we say that the sets are **disjoint**.

Definition 2.11 (**Disjoint**). *Two sets A and B are* **disjoint** *if*

$$A \cap B = \emptyset. \tag{2.12}$$

For a collection of sets $\{A_1, A_2, \ldots, A_n\}$, we say that the collection is disjoint if, for any pair $i \neq j$,

$$A_i \cap A_j = \emptyset. \tag{2.13}$$

A pictorial interpretation can be found in **Figure 2.9**.

Example 2.15(a). Let $A = \{x > 1\}$ and $B = \{x < 0\}$. Then A and B are disjoint.

Example 2.15(b). Let $A = \{1, 2, 3\}$ and $B = \emptyset$. Then A and B are disjoint.

Example 2.15(c). Let $A = (0, 1)$ and $B = [1, 2)$. Then A and B are disjoint.

With the definition of disjoint, we can now define the powerful concept of **partition**.

Definition 2.12 (**Partition**). *A collection of sets $\{A_1, \ldots, A_n\}$ is a* **partition** *of the universal set Ω if it satisfies the following conditions:*

- *(***non-overlap***) $\{A_1, \ldots, A_n\}$ is disjoint:*

$$A_i \cap A_j = \emptyset. \tag{2.14}$$

- *(***decompose***) Union of $\{A_1, \ldots, A_n\}$ gives the universal set:*

$$\bigcup_{i=1}^{n} A_i = \Omega. \tag{2.15}$$

In plain language, a partition is a collection of non-overlapping subsets whose union is the universal set. Partition is important because it is a **decomposition** of Ω into a smaller subset, and since these subsets do not overlap, they can be analyzed separately. Partition is a handy tool for studying probability because it allows us to decouple complex events by treating them as isolated sub-events.

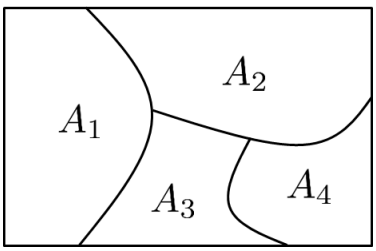


Figure 2.10: A partition of Ω contains disjoint subsets of which the union gives us Ω.

Example 2.16. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. The following sets form a partition:

$$A_1 = \{1, 2, 3\}, \quad A_2 = \{4, 5\}, \quad A_3 = \{6\}$$

Example 2.17. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. The collection

$$A_1 = \{1, 2, 3\}, \quad A_2 = \{4, 5\}, \quad A_3 = \{5, 6\}$$

does not form a partition, because $A_2 \cap A_3 = \{5\}$.

If $\{A_1, A_2, \ldots, A_n\}$ forms a partition of the universal set Ω, then for any $B \subseteq \Omega$, we can decompose B into n disjoint subsets: $B \cap A_1$, $B \cap A_2$, $\ldots B \cap A_n$. Two properties hold:

- $B \cap A_i$ and $B \cap A_j$ are disjoint if $i \neq j$.
- The union of $B \cap A_1$, $B \cap A_2$, $\ldots B \cap A_n$ is B.

Practice Exercise 2.7. Prove the above two statements.

Solution. To prove the first statement, we can pick $\xi \in (B \cap A_i)$. This means that $\xi \in B$ and $\xi \in A_i$. Since $\xi \in A_i$, it cannot be in A_j because A_i and A_j are disjoint. Therefore ξ cannot live in $B \cap A_j$. This completes the proof, because we just showed that any $\xi \in B \cap A_i$ cannot simultaneously live in $B \cap A_j$.

To prove the second statement, we pick $\xi \in \bigcup_{i=1}^{n}(B \cap A_i)$. Since ξ lives in the union, it has to live in at least one of the $(B \cap A_i)$ for some i. Now suppose $\xi \in B \cap A_i$. This means that ξ is in both B and A_i, so it must live in B. Therefore, $\bigcup_{i=1}^{n}(B \cap A_i) \subseteq B$. Now, suppose we pick $\xi \in B$. Then since it is an element in B, it must be an element in all of the $(B \cap A_i)$'s for any i. Therefore, $\xi \in \bigcup_{i=1}^{n}(B \cap A_i)$, and so we showed that $B \subseteq \bigcup_{i=1}^{n}(B \cap A_i)$. Combining the two directions, we conclude that $\bigcup_{i=1}^{n}(B \cap A_i) = B$.

Example 2.18. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and let a partition of Ω be $A_1 = \{1, 2, 3\}$, $A_2 = \{4, 5\}$, $A_3 = \{6\}$. Let $B = \{1, 3, 4\}$. Then, by the result we just proved, B can be decomposed into three subsets:

$$B \cap A_1 = \{1, 3\}, \quad B \cap A_2 = \{4\}, \quad B \cap A_3 = \emptyset.$$

Thus we can see that $B \cap A_1$, $B \cap A_2$ and $B \cap A_3$ are disjoint. Furthermore, the union of these three sets gives B.

2.1.9 Set operations

When handling multiple sets, it would be useful to have some basic set operations. There are four basic theorems concerning set operations that you need to know for our purposes in this book:

Theorem 2.2 (**Commutative**). *(Order does not matter)*

$$A \cap B = B \cap A, \quad \textit{and} \quad A \cup B = B \cup A. \tag{2.16}$$

Theorem 2.3 (**Associative**). *(How to do multiple union and intersection)*

$$\begin{aligned} A \cup (B \cup C) &= (A \cup B) \cup C, \\ A \cap (B \cap C) &= (A \cap B) \cap C. \end{aligned} \tag{2.17}$$

Theorem 2.4 (**Distributive**). *(How to mix union and intersection)*

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C). \tag{2.18}$$

Theorem 2.5 (**De Morgan's Law**). *(How to complement over intersection and union)*

$$(A \cap B)^c = A^c \cup B^c,$$
$$(A \cup B)^c = A^c \cap B^c. \tag{2.19}$$

Example 2.19. Consider $[1,4] \cap ([0,2] \cup [3,5])$. By the distributive property we can simplify the set as

$$\begin{aligned}[1,4] \cap ([0,2] \cup [3,5]) &= ([1,4] \cap [0,2]) \cup ([1,4] \cap [3,5]) \\ &= [1,2] \cup [3,4].\end{aligned}$$

Example 2.20. Consider $([0,1] \cup [2,3])^c$. By De Morgan's Law we can rewrite the set as

$$([0,2] \cup [1,3])^c = [0,2]^c \cap [1,3]^c.$$

2.1.10 Closing remarks about set theory

It should be apparent why set theory is useful: it shows us how to combine, split, and remove sets. In **Figure 2.11** we depict the intersection of two sets $A = \{\text{even number}\}$ and $B = \{\text{less than or equal to 3}\}$. Set theory tells us how to define the intersection so that the probability can be applied to the resulting set.

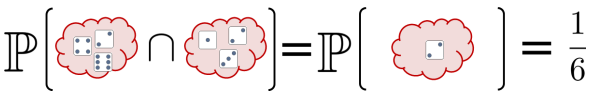


Figure 2.11: When there are two events A and B, the probability of $A \cap B$ is determined by first taking the intersection of the two sets and then evaluating its probability.

Universal sets and empty sets are useful too. Universal sets cover all the possible outcomes of an experiment, so we should expect $\mathbb{P}[\Omega] = 1$. Empty sets contain nothing, and so we should expect $\mathbb{P}[\emptyset] = 0$. These two properties are essential to define a probability because no probability can be greater than 1, and no probability can be less than 0.

2.2 Probability Space

We now formally define probability. Our discussion will be based on the slogan **probability is a measure of the size of a set**. Three elements constitute a **probability space**:

- **Sample Space** Ω: The set of all possible outcomes from an experiment.
- **Event Space** $\mathcal{F}$: The collection of all possible events. An event E is a subset in Ω that defines an outcome or a combination of outcomes.
- **Probability Law** $\mathbb{P}$: A mapping from an event E to a number $\mathbb{P}[E]$ which, ideally, measures the size of the event.

Therefore, whenever you talk about "probability," you need to specify the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ to define the probability space.

The necessity of the three elements is illustrated in **Figure 2.12**. The **sample space** is the interface with the **physical world**. It is the collection of all possible states that can result from an experiment. Some outcomes are more likely to happen, and some are less likely, but this does not matter because the sample space contains every possible outcome. The **probability law** is the interface with the **data analysis**. It is this law that defines the likelihood of each of the outcomes. However, since the probability law measures the size of a set, the probability law itself must be a function, a function whose argument is a set and whose value is a number. An outcome in the sample space is not a set. Instead, a subset in the sample space is a set. Therefore, the probability should input a subset and map it to a number. The collection of all possible subsets is the **event space**.

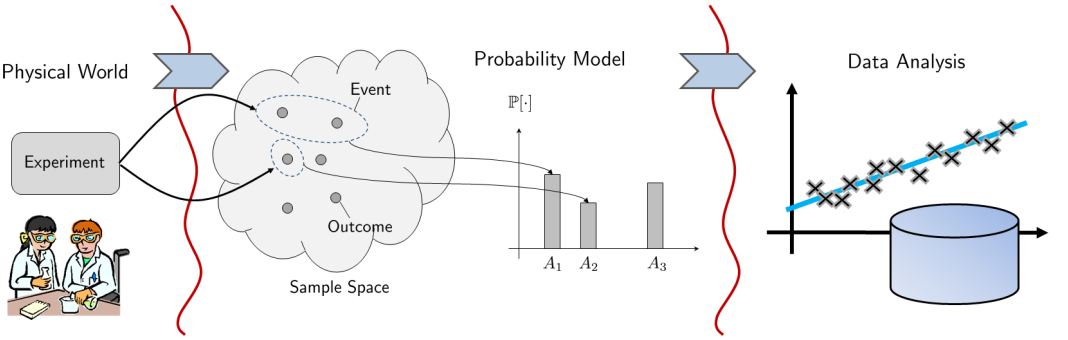


Figure 2.12: Given an experiment, we define the collection of all outcomes as the sample space. A subset in the sample space is called an event. The probability law is a mapping that maps an event to a number that denotes the size of the event.

A perceptive reader like you may be wondering why we want to complicate things to this degree when calculating probability is trivial, e.g., throwing a die gives us a probability $\frac{1}{6}$ per face. In a simple world where problems are that easy, you can surely ignore all these complications and proceed to the answer $\frac{1}{6}$. However, modern data analysis is not so easy. If we are given an image of size 64×64 pixels, how do we tell whether this image is of a cat or a dog? We need to construct a probability model that tells us the likelihood of having a

particular 64×64 image. What should be included in this probability model? We need to know all the possible cases (**the sample space**), all the possible events (**the event space**), and the probability of each of the events (**the probability law**). If we know all these, then our decision will be theoretically optimal. Of course, for high-dimensional data like images, we need approximations to such a probability model. However, we first need to understand the theoretical foundation of the probability space to know what approximations would make sense.

2.2.1 Sample space Ω

We start by defining the sample space Ω. Given an experiment, the **sample space** Ω is the set containing all possible outcomes of the experiment.

Definition 2.13. *A* **sample space** Ω *is the set of all possible outcomes from an experiment. We denote* ξ *as an element in* Ω.

A sample space can contain discrete outcomes or continuous outcomes, as shown in the examples below and **Figure 2.13**.

Example 2.21: (Discrete Outcomes)

- Coin flip: $\Omega = \{H,\ T\}$.
- Throw a die: $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$.
- Paper / scissor / stone: $\Omega = \{\text{paper}, \text{scissor}, \text{stone}\}$.
- Draw an even integer: $\Omega = \{2, 4, 6, 8, \ldots\}$.

Example 2.22: (Continuous Outcomes)

- Waiting time for a bus in West Lafayette: $\Omega = \{t \mid 0 \leq t \leq 30 \text{ minutes}\}$.
- Phase angle of a voltage: $\Omega = \{\theta \mid 0 \leq \theta \leq 2\pi\}$.
- Frequency of a pitch: $\Omega = \{f \mid 0 \leq f \leq f_{\max}\}$.

Figure 2.13 also shows a **functional** example of the sample space. In this case, the sample space contains **functions**. For example,

- Set of all straight lines in 2D:

$$\Omega = \{f \mid f(x) = ax + b,\ a, b \in \mathbb{R}\}.$$

- Set of all cosine functions with a phase offset:

$$\Omega = \{f \mid f(t) = \cos(2\pi\omega_0 t + \Theta),\ 0 \leq \Theta \leq 2\pi\}.$$

As we see from the above examples, the sample space is nothing but a universal set. The elements inside the sample space are the outcomes of the experiment. If you change

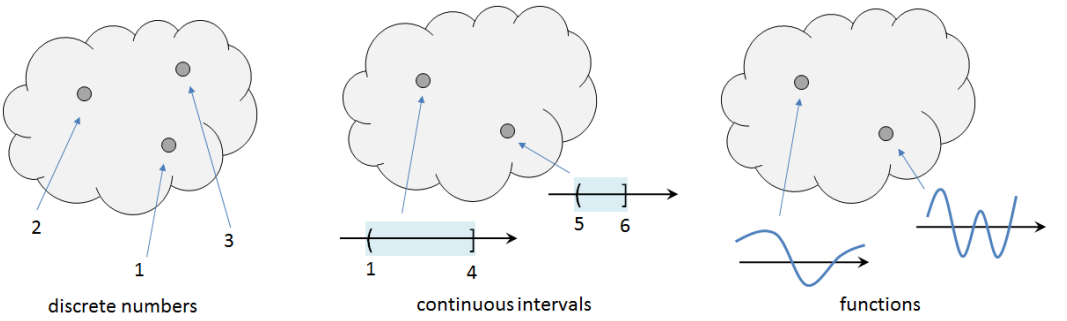


Figure 2.13: The sample space can take various forms: it can contain discrete numbers, or continuous intervals, or even functions.

the experiment, the possible outcomes will be different so that the sample space will be different. For example, flipping a coin has different possible outcomes from throwing a die.

What if we want to describe a composite experiment where we flip a coin and throw a die? Here is the sample space:

Example 2.23: If the experiment contains flipping a coin and throwing a die, then the sample space is

$$\Big\{(H,\boxdot),(H,\boxdot),(H,\boxdot),(H,\boxdot),(H,\boxdot),(H,\boxdot),$$
$$(T,\boxdot),(T,\boxdot),(T,\boxdot),(T,\boxdot),(T,\boxdot),(T,\boxdot)\Big\}.$$

In this sample space, each element is a pair of outcomes.

Practice Exercise 2.8. There are 8 processors on a computer. A computer job scheduler chooses one processor randomly. What is the sample space? If the computer job scheduler can choose two processors at once, what is the sample space then?

Solution. The sample space of the first case is $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$. The sample space of the second case is $\Omega = \{(1, 2), (1, 3), (1, 4), \ldots, (7, 8)\}$.

Practice Exercise 2.9. A cell phone tower has a circular average coverage area of radius of 10 km. We observe the source locations of calls received by the tower. What is the sample space of all possible source locations?

Solution. Assume that the center of the tower is located at (x_0, y_0). The sample space is the set

$$\Omega = \{(x, y) \mid \sqrt{(x - x_0)^2 + (y - y_0)^2} \leq 10\}.$$

Not every set can be a sample space. A sample space must be **exhaustive** and **exclusive**. The term "exhaustive" means that the sample space has to cover **all** possible outcomes. If

there is one possible outcome that is left out, then the set is no longer a sample space. The term "exclusive" means that the sample space contains unique elements so that there is no repetition of elements.

Example 2.24. (Counterexamples)

The following two examples are NOT sample spaces.

- Throw a die: $\Omega = \{1, 2, 3\}$ is not a sample space because it is not **exhaustive**.
- Throw a die: $\Omega = \{1, 1, 2, 3, 4, 5, 6\}$ is not a sample space because it is not **exclusive**.

Therefore, a valid sample space must contain all possible outcomes, and each element must be unique.

We summarize the concept of a sample space as follows.

What is a sample space Ω?

- A sample space Ω is the collection of all possible outcomes.
- The outcomes can be numbers, alphabets, vectors, or functions. The outcomes can also be images, videos, EEG signals, audio speeches, etc.
- Ω must be exhaustive and exclusive.

2.2.2 Event space $\mathcal{F}$

The sample space contains all the possible outcomes. However, in many practical situations, we are not interested in each of the individual outcomes; we are interested in the *combinations* of the outcomes. For example, when throwing a die, we may ask "What is the probability of rolling an odd number?" or "What is the probability of rolling a number that is less than 3?" Clearly, "odd number" is not an outcome of the experiment because the possible outcomes are $\{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. We call "odd number" an **event**. An event must be a subset in the sample space.

Definition 2.14. *An* **event** E *is a subset in the sample space* Ω*. The set of all possible events is denoted as* $\mathcal{F}$*.*

While this definition is extremely simple, we need to keep in mind a few facts about events. First, an outcome ξ is an element in Ω but an event E is a subset contained in Ω, i.e., $E \subseteq \Omega$. Thus, an event can contain one outcome but it can also contain many outcomes. The following example shows a few cases of events:

Example 2.25. Throw a die. Let $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. The following are two possible events, as illustrated in **Figure 2.14**.

- $E_1 = \{\text{even numbers}\} = \{\boxdot, \boxdot, \boxdot\}$.

- $E_2 = \{\text{less than } 3\} = \{\boxdot, \boxdot\}$.

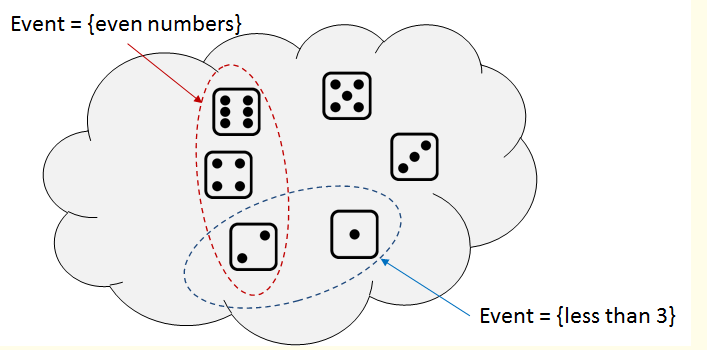


Figure 2.14: Two examples of events: The first event contains numbers $\{2, 4, 6\}$, and the second event contains numbers $\{1, 2\}$.

Practice Exercise 2.10. The "ping" command is used to measure round-trip times for Internet packets. What is the sample space of all possible round-trip times? What is the event that a round-trip time is between 10 ms and 20 ms?

Solution. The sample space is $\Omega = [0, \infty)$. The event is $E = [10, 20]$.

Practice Exercise 2.11. A cell phone tower has a circular average coverage area of radius 10 km. We observe the source locations of calls received by the tower. What is the event when the source location of a call is between 2 km and 5 km from the tower?

Solution. Assume that the center of the tower is located at (x_0, y_0). The event is $E = \{(x, y) \mid 2 \leq \sqrt{(x - x_0)^2 + (y - y_0)^2} \leq 5\}$.

The second point we should remember is the cardinality of Ω and that of $\mathcal{F}$. A sample space containing n elements has a cardinality n. However, the event space constructed from Ω will contain 2^n events. To see why this is so, let's consider the following example.

Example 2.26. Consider an experiment with 3 outcomes $\Omega = \{\clubsuit, \heartsuit, \maltese\}$. We can list out all the possible events: $\emptyset$, $\{\clubsuit\}$, $\{\heartsuit\}$, $\{\maltese\}$, $\{\clubsuit, \heartsuit\}$, $\{\clubsuit, \maltese\}$, $\{\heartsuit, \clubsuit\}$, $\{\clubsuit, \heartsuit, \maltese\}$. So in total there are $2^3 = 8$ possible events. **Figure 2.15** depicts the situation. What is the difference between $\clubsuit$ and $\{\clubsuit\}$? The former is an element, whereas the latter is a set. Thus, $\{\clubsuit\}$ is an event but $\clubsuit$ is not an event. Why is $\emptyset$ an event? Because we can ask "What is the probability that we get an odd number and an even number?" The probability is obviously zero, but the reason it is zero is that the event is an empty set.

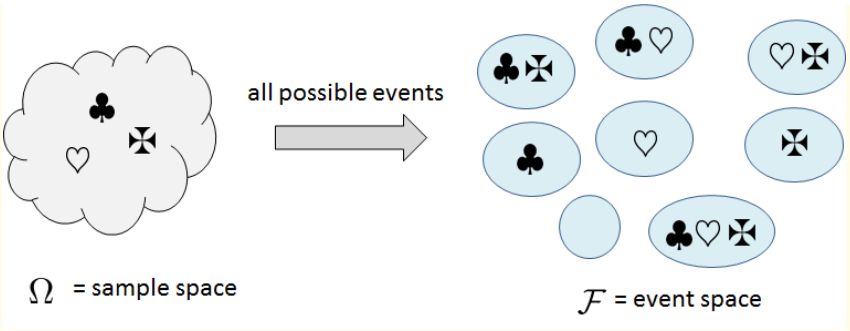


Figure 2.15: The event space contains all the possible subsets inside the sample space.

In general, if there are n elements in the sample space, then the number of events is 2^n. To see why this is true, we can assign to each element a binary value: either 0 or 1. For example, in Table 2.1 we consider throwing a die. For each of the six faces, we assign a binary code. This will give us a binary string for each event. For example, the event $\{\boxdot, \boxtimes\}$ is encoded as the binary string 100010 because only $\boxdot$ and $\boxtimes$ are activated. We can count the total number of unique strings, which is the number of strings that can be constructed from n bits. It is easily seen that this number is 2^n.

Event	$\boxdot$	$\boxdot$	$\boxdot$	$\boxdot$	$\boxdot$	$\boxdot$	Binary Code
$\emptyset$	×	×	×	×	×	×	000000
$\{\boxdot, \boxtimes\}$	◯	×	×	×	◯	×	100010
$\{\boxdot, \boxdot, \boxtimes\}$	×	×	◯	◯	◯	×	001110
⋮		⋮		⋮			⋮
$\{\boxdot, \boxdot, \boxdot, \boxtimes, \boxdot\}$	×	◯	◯	◯	◯	◯	011111
$\{\boxdot, \boxdot, \boxdot, \boxdot, \boxtimes, \boxdot\}$	◯	◯	◯	◯	◯	◯	111111

Table 2.1: An event space contains 2^n events, where n is the number of elements in the sample space. To see this, we encode each outcome with a binary code. The resulting binary string then forms a unique index of the event. Counting the total number of events gives us the cardinality of the event space.

The box below summarizes what you need to know about event spaces.

What is an event space $\mathcal{F}$?

- An event space $\mathcal{F}$ is the set of all possible subsets. It is a set of sets.
- We need $\mathcal{F}$ because the probability law $\mathbb{P}$ is mapping a set to a number. $\mathbb{P}$ does not take an outcome from Ω but a subset inside Ω.

Event spaces: Some advanced topics

The following discussions can be skipped if it is your first time reading the book.

What else do we need to take care of in order to ensure that an event is well defined? A few set operations seem to be necessary. For example, if $E_1 = \{\boxdot\}$ and $E_2 = \{\boxdot\}$ are events, it is necessary that $E = E_1 \cup E_2 = \{\boxdot, \boxdot\}$ is an event too. Another example: if $E_1 = \{\boxtimes, \boxplus\}$ and $E_2 = \{\boxdot, \boxtimes\}$ are events, then it is necessary that $E = E_1 \cap E_2 = \{\boxtimes\}$ is also an event. The third example: if $E_1 = \{\boxdot, \boxdot, \boxtimes, \boxplus\}$ is an event, then $E = E_1^c = \{\boxdot, \boxdot\}$ should be an event. As you can see, there is nothing sophisticated in these examples. They are just some basic set operations. We want to ensure that the event space is **closed** under these set operations. That is, we do not want to be surprised by finding that a set constructed from two events is not an event. However, since all set operations can be constructed from union, intersection and complement, ensuring that the event space is closed under these three operations effectively ensures that it is closed to **all** set operations.

The formal way to guarantee these is the notion of a **field**. This term may seem to be abstract, but it is indeed quite useful:

Definition 2.15. *For an event space $\mathcal{F}$ to be valid, $\mathcal{F}$ must be a* **field** *$\mathcal{F}$. It is a field if it satisfies the following conditions*

- *$\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.*
- *(Closed under complement) If $F \in \mathcal{F}$, then also $F^c \in \mathcal{F}$.*
- *(Closed under union and intersection) If $F_1 \in \mathcal{F}$ and $F_2 \in \mathcal{F}$, then $F_1 \cap F_2 \in \mathcal{F}$ and $F_1 \cup F_2 \in \mathcal{F}$.*

For a finite set, i.e., a set that contains n elements, the collection of all possible subsets is indeed a field. This is not difficult to see if you consider rolling a die. For example, if $E = \{\boxdot, \boxdot, \boxtimes, \boxplus\}$ is inside $\mathcal{F}$, then $E^c = \{\boxdot, \boxdot\}$ is also inside $\mathcal{F}$. This is because $\mathcal{F}$ consists of 2^n subsets each being encoded by a unique binary string. So if $E = 001111$, then $E^c = 110000$, which is also in $\mathcal{F}$. Similar reasoning applies to intersection and union.

At this point, you may ask:

- **Why bother constructing a field?** The answer is that probability is a measure of the size of a set, so we must input a set to a probability measure $\mathbb{P}$ to get a number. The set being input to $\mathbb{P}$ must be a subset inside the sample space; otherwise, it will be undefined. If we regard $\mathbb{P}$ as a mapping, we need to specify the collection of all its inputs, which is the set of all subsets, i.e., the event space. So if we do not define the field, there is no way to define the measure $\mathbb{P}$.
- **What if the event space is not a field?** If the event space is not a field, then we can easily construct pathological cases where we cannot assign a probability. For example, if the event space is not a field, then it would be possible that the complement of $E = \{\boxdot, \boxdot, \boxtimes, \boxplus\}$ (which is $E^c = \{\boxdot, \boxdot\}$) is not an event. This just does not make sense.

The concept of a field is sufficient for finite sample spaces. However, there are two other types of sample spaces where the concept of a field is inadequate. The first type of

sets consists of the **countably infinite** sets, and the second type consists of the sets defined on the **real line**. There are other types of sets, but these two have important practical applications. Therefore, we need to have a basic understanding of these two types.

Sigma-field

The difficulty of a countably infinite set is that there are infinitely many subsets in the field of a countably infinite set. Having a finite union and a finite intersection is insufficient to ensure the closedness of all intersections and unions. In particular, having $F_1 \cup F_2 \in \mathcal{F}$ does not automatically give us $\bigcup_{n=1}^{\infty} F_n \in \mathcal{F}$ because the latter is an infinite union. Therefore, for countably infinite sets, their requirements to be a field are more restrictive as we need to ensure infinite intersection and union. The resulting field is called the σ-field.

Definition 2.16. *A sigma-field (σ-**field**) $\mathcal{F}$ is a field such that*

- *$\mathcal{F}$ is a field, and*
- *if $F_1, F_2, \ldots \in \mathcal{F}$, then the union $\bigcup_{i=1}^{\infty} F_i$ and the intersection $\bigcap_{i=1}^{\infty} F_i$ are both in $\mathcal{F}$.*

When do we need a σ-field? When the sample space is countable and has infinitely many elements. For example, if the sample space contains all integers, then the collection of all possible subsets is a σ-field. For another, if $E_1 = \{2\}$, $E_2 = \{4\}$, $E_3 = \{6\}$, $\ldots$, then $\bigcup_{n=1}^{\infty} E_n = \{2, 4, 6, 8, \ldots\} = \{\text{positive even numbers}\}$. Clearly, we want $\bigcup_{n=1}^{\infty} E_n$ to live in the sample space.

Borel sigma-field

While a sigma-field allows us to consider countable sets of events, it is still insufficient for considering events defined on the real line, e.g., time, as these events are not countable. So how do we define an event on the real line? It turns out that we need a different way to define the **smallest unit**. For finite sets and countable sets, the smallest units are the elements themselves because we can **count** them. For the real line, we cannot count the elements because any non-empty interval is uncountably infinite.

The smallest unit we use to construct a field for the real line is a semi-closed interval

$$(-\infty, b] \stackrel{\text{def}}{=} \{x \mid -\infty < x \leq b\}.$$

The **Borel σ-field** is defined as the sigma-field generated by the semi-closed intervals.

Definition 2.17. *The **Borel σ-field** $\mathcal{B}$ is a σ-field generated from semi-closed intervals:*

$$(-\infty, b] \stackrel{\text{def}}{=} \{x \mid -\infty < x \leq b\}.$$

The difference between the Borel σ-field $\mathcal{B}$ and a regular σ-field is how we measure the subsets. In a σ-field, we count the elements in the subsets, whereas, in a Borel σ-field, we use the semi-closed intervals to measure the subsets.

Being a field, the Borel σ-field is closed under complement, union, and intersection. In particular, subsets of the following forms are also in the Borel σ-field $\mathcal{B}$:

$$(a,b),\ [a,b],\ (a,b],\ [a,b),\ [a,\infty),\ (a,\infty),\ (-\infty,b],\ \{b\}.$$

For example, (a,∞) can be constructed from $(-\infty,a]^c$, and $(a,b]$ can be constructed by taking the intersection of $(-\infty,b]$ and (a,∞).

Example 2.27: Waiting for a bus. Let $\Omega = \{0 \le t \le 30\}$. The Borel σ-field contains all semi-closed intervals $(a,b]$, where $0 \le a \le b \le 30$. Here are two possible events:

- $F_1 = \{\text{less than 10 minutes}\} = \{0 \le t < 10\} = \{0\} \cup (\{0 < t \le 10\} \cap \{10\}^c)$.
- $F_2 = \{\text{more than 20 minutes}\} = \{20 < t \le 30\}$.

Further discussion of the Borel σ-field can be found in Leon-Garcia (3rd Edition,) Chapter 2.9.

This is the end of the discussion. Please join us again.

2.2.3 Probability law $\mathbb{P}$

The third component of a probability space is the probability law $\mathbb{P}$. Its job is to assign a number to an event.

Definition 2.18. *A* **probability law** *is a function* $\mathbb{P} : \mathcal{F} \to [0,1]$ *of an event* E *to a real number in* $[0, 1]$.

The probability law is thus a **function**, and therefore we must specify the input and the output. The input to $\mathbb{P}$ is an event E, which is a subset in Ω and an element in $\mathcal{F}$. The output of $\mathbb{P}$ is a number between 0 and 1, which we call the **probability**.

The definition above does not specify how an event is being mapped to a number. However, since probability is a measure of the size of a set, a meaningful $\mathbb{P}$ should be **consistent** for all events in $\mathcal{F}$. This requires some rules, known as the **axioms of probability**, when we define the $\mathbb{P}$. Any probability law $\mathbb{P}$ must satisfy these axioms; otherwise, we will see contradictions. We will discuss the axioms in the next section. For now, let us look at two examples to make sure we understand the functional nature of $\mathbb{P}$.

Example 2.28. Consider flipping a coin. The event space is $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$. We can define the probability law as

$$\mathbb{P}[\emptyset] = 0,\ \ \mathbb{P}[\{H\}] = \frac{1}{2},\ \ \mathbb{P}[\{T\}] = \frac{1}{2},\ \ \mathbb{P}[\Omega] = 1,$$

as shown in **Figure 2.16**. This $\mathbb{P}$ is clearly consistent for all the events in $\mathcal{F}$.

Is it possible to construct an invalid $\mathbb{P}$? Certainly. Consider the following proba-

bility law:

$$\mathbb{P}[\emptyset] = 0, \quad \mathbb{P}[\{H\}] = \frac{1}{3}, \quad \mathbb{P}[\{T\}] = \frac{1}{3}, \quad \mathbb{P}[\Omega] = 1.$$

This law is invalid because the individual events are $\mathbb{P}[\{H\}] = \frac{1}{3}$ and $\mathbb{P}[\{T\}] = \frac{1}{3}$ but the union is $\mathbb{P}[\Omega] = 1$. To fix this problem, one possible solution is to define the probability law as

$$\mathbb{P}[\emptyset] = 0, \quad \mathbb{P}[\{H\}] = \frac{1}{3}, \quad \mathbb{P}[\{T\}] = \frac{2}{3}, \quad \mathbb{P}[\Omega] = 1.$$

Then, the probabilities for all the events are well defined and consistent.

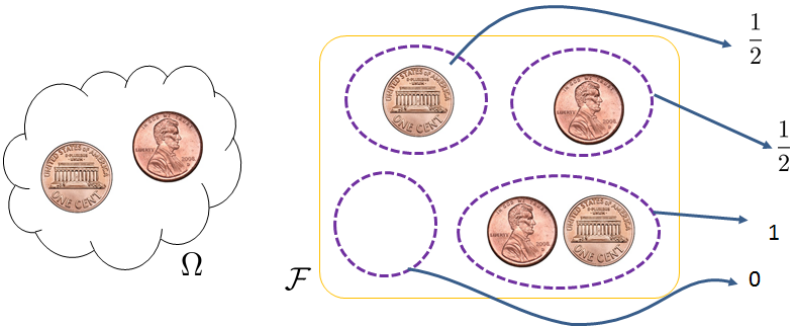


Figure 2.16: A probability law is a mapping from an event to a number. A probability law cannot be arbitrarily assigned; it must satisfy the axioms of probability.

Example 2.29. Consider a sample space containing three elements $\Omega = \{\clubsuit, \heartsuit, \maltese\}$. The event space is then $\mathcal{F} = \left\{\emptyset, \{\clubsuit\}, \{\heartsuit\}, \{\maltese\}, \{\clubsuit, \heartsuit\}, \{\heartsuit, \maltese\}, \{\clubsuit, \maltese\}, \{\clubsuit, \heartsuit, \maltese\}\right\}$. One possible $\mathbb{P}$ we could define would be

$$\mathbb{P}[\emptyset] = 0, \quad \mathbb{P}[\{\clubsuit\}] = \mathbb{P}[\{\heartsuit\}] = \mathbb{P}[\{\maltese\}] = \frac{1}{3},$$
$$\mathbb{P}[\{\clubsuit, \heartsuit\}] = \mathbb{P}[\{\clubsuit, \maltese\}] = \mathbb{P}[\{\heartsuit, \maltese\}] = \frac{2}{3}, \quad \mathbb{P}[\{\clubsuit, \heartsuit, \maltese\}] = 1.$$

What is a probability law $\mathbb{P}$?

- A probability law $\mathbb{P}$ is a **function**.
- It takes a subset (an element in $\mathcal{F}$) and maps it to a number between 0 and 1.
- $\mathbb{P}$ is a **measure** of the size of a set.
- For $\mathbb{P}$ to be valid, it must satisfy the **axioms of probability**.

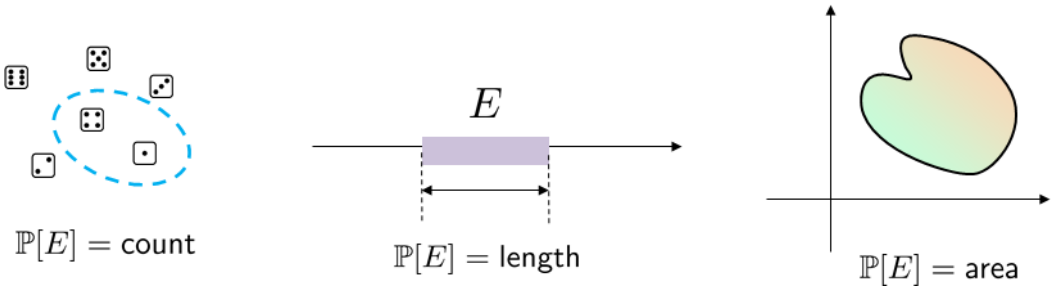


Figure 2.17: Probability is a measure of the size of a set. The probability can be a counter that counts the number of elements, a ruler that measures the length of an interval, or an integration that measures the area of a region.

A probability law $\mathbb{P}$ is a measure

Consider the word "measure" in our slogan: **probability is a measure of the size of a set**. Depending on the nature of the set, the measure can be a counter, ruler, scale, or even a stopwatch. So far, all the examples we have seen are based on sets with a finite number of elements. For these sets, the natural choice of the probability measure is a counter. However, if the sets are intervals on the real line or regions in a plane, we need a different probability law to measure their size. Let's look at the examples shown in **Figure 2.17**.

Example 2.30 (Finite Set). Consider throwing a die, so that

$$\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}.$$

Then the probability measure is a counter that reports the number of elements. If the die is fair, i.e., all the 6 faces have equal probability of happening, then an event $E = \{\boxdot, \boxdot\}$ will have a probability $\mathbb{P}[E] = \frac{2}{6}$.

Example 2.31 (Intervals). Suppose that the sample space is a unit interval $\Omega = [0, 1]$. Let E be an event such that $E = [a, b]$ where a, b are numbers in $[0, 1]$. Then the probability measure is a ruler that measures the length of the intervals. If all the numbers on the real line have equal probability of appearing, then $\mathbb{P}[E] = b - a$.

Example 2.32 (Regions). Suppose that the sample space is the square $\Omega = [-1, 1] \times [-1, 1]$. Let E be a circle such that $E = \{(x, y) | x^2 + y^2 < r^2\}$, where $r < 1$. Then the probability measure is an area measure that returns us the area of E. If we assume that all coordinates in Ω are equally probable, then $\mathbb{P}[E] = \pi r^2$, for $r < 1$.

Because probability is a measure of the size of a set, two sets can be compared according to their probability measures. For example, if $\Omega = \{\clubsuit, \heartsuit, \maltese\}$, and if $E_1 = \{\clubsuit\}$ and $E_2 = \{\clubsuit, \heartsuit\}$, then one possible $\mathbb{P}$ is to assign $\mathbb{P}[E_1] = \mathbb{P}[\{\clubsuit\}] = \frac{1}{3}$ and $\mathbb{P}[E_2] = \mathbb{P}[\{\clubsuit, \heartsuit\}] = 2/3$.

In this particular case, we see that $E_1 \subseteq E_2$ and thus

$$\mathbb{P}[E_1] \leq \mathbb{P}[E_2].$$

Let's now consider the term "size." Notice that the concept of the size of a set is not limited to the number of elements. A better way to think about size is to imagine that it is the weight of the set. This might may seem fanciful at first, but it is quite natural. Consider the following example.

Example 2.33. (**Discrete events with different weights**) Suppose we have a sample space $\Omega = \{\clubsuit, \heartsuit, ✠\}$. Let us assign a different probability to each outcome:

$$\mathbb{P}[\{\clubsuit\}] = \frac{2}{6}, \quad \mathbb{P}[\{\heartsuit\}] = \frac{1}{6}, \quad \mathbb{P}[\{✠\}] = \frac{3}{6}.$$

As illustrated in **Figure 2.18**, since each outcome has a different weight, when determining the probability of a set of outcomes we can add these weights (instead of counting the number of outcomes). For example, when reporting $\mathbb{P}[\{\clubsuit\}]$ we find its weight $\mathbb{P}[\{\clubsuit\}] = \frac{2}{6}$, whereas when reporting $\mathbb{P}[\{\heartsuit, ✠\}]$ we find the sum of their weights $\mathbb{P}[\{\heartsuit, ✠\}] = \frac{1}{6} + \frac{3}{6} = \frac{4}{6}$. Therefore, the notion of size does not refer to the number of elements but to the total weight of these elements.

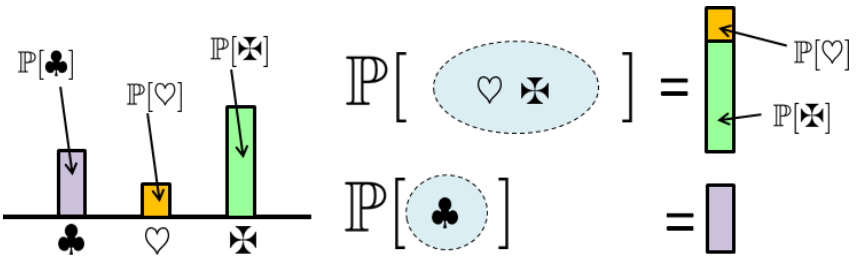


Figure 2.18: This example shows the "weights" of three elements in a set. The weights are numbers between 0 and 1 such that the sum is 1. When applying a probability measure to this set, we sum the weights for the elements in the events being considered. For example, $\mathbb{P}[\heartsuit, ✠]$ = yellow + green, and $\mathbb{P}[\clubsuit]$ = purple.

Example 2.34. (**Continuous events with different weights**) Suppose that the sample space is an interval, say $\Omega = [-1, 1]$. On this interval we define a weighting function $f(x)$ where $f(x_0)$ specifies the weight for x_0. Because Ω is an interval, events defined on this Ω must also be intervals. For example, we can consider two events $E_1 = [a, b]$ and $E_2 = [c, d]$. The probabilities of these events are $\mathbb{P}[E_1] = \int_a^b f(x)\, dx$ and $\mathbb{P}[E_2] = \int_c^d f(x)\, dx$, as shown in **Figure 2.19**.

Viewing probability as a measure is not just a game for mathematicians; rather, it has fundamental significance for several reasons. First, it eliminates any dependency on probability as relative frequency from the frequentist point of view. Relative frequency is a

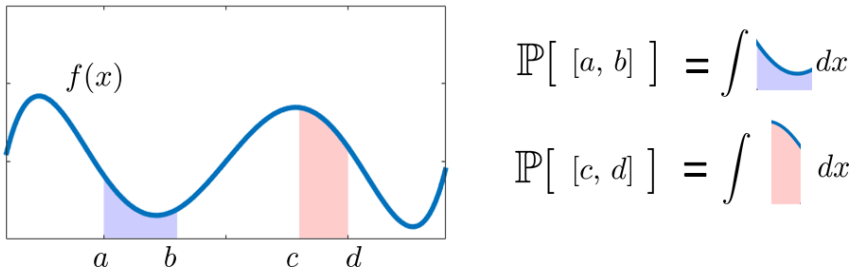


Figure 2.19: If the sample space is an interval on the real line, then the probability of an event is the area under the curve of the weighting function.

narrowly defined concept that is largely limited to discrete events, e.g., flipping a coin. While we can assign weights to coin-toss events to deal with those biased coins, the extension to continuous events becomes problematic. By thinking of probability as a measure, we can generalize the notion to apply to intervals, areas, volumes, and so on.

Second, viewing probability as a measure forces us to disentangle an **event** from **measures**. An event is a subset in the sample space. It has nothing to do with the measure (e.g., a ruler) you use to measure the event. The measure, on the other hand, specifies the weighting function you apply to measure the event when computing the probability. For example, let $\Omega = [-1, 1]$ be an interval, and let $E = [a, b]$ be an event. We can define two weighting functions $f(x)$ and $g(x)$. Correspondingly, we will have two different probability measures $\mathbb{F}$ and $\mathbb{G}$ such that

$$\begin{aligned} \mathbb{F}([a,b]) &= \int_E d\mathbb{F} = \int_a^b f(x)\, dx, \\ \mathbb{G}([a,b]) &= \int_E d\mathbb{G} = \int_a^b g(x)\, dx. \end{aligned} \tag{2.20}$$

To make sense of these notations, consider only $\mathbb{P}[[a, b]]$ and not $\mathbb{F}([a, b])$ and $\mathbb{G}([a, b])$. As you can see, the event for both measures is $E = [a, b]$ but the measures are different. Therefore, the values of the probability are different.

Example 2.35. (**Two probability laws are different if their weighting functions are different.**) Consider two different weighting functions for throwing a die. The first one assigns probability as the following:

$$\begin{aligned} \mathbb{P}[\{\boxdot\}] = \frac{1}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{3}{12}, \\ \mathbb{P}[\{\boxdot\}] = \frac{4}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{1}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{1}{12}, \end{aligned}$$

whereas the second function assigns the probability like this:

$$\begin{aligned} \mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12}, \\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12}. \end{aligned}$$

Let an event $E = \{\boxdot, \boxdot\}$. Let $\mathbb{F}$ be the measure using the first set of probabilities, and let $\mathbb{G}$ be the measure of the second set of probabilities. Then,

$$\mathbb{F}(E) = \mathbb{F}(\{\boxdot, \boxdot\}) = \frac{1}{12} + \frac{2}{12} = \frac{3}{12},$$
$$\mathbb{G}(E) = \mathbb{G}(\{\boxdot, \boxdot\}) = \frac{2}{12} + \frac{2}{12} = \frac{4}{12}.$$

Therefore, although the events are the same, the two different measures will give us two different probability values.

Remark. The notation $\int_E d\mathbb{F}$ in Equation (2.20) is known as the **Lebesgue integral**. You should be aware of this notation, but the theory of Lebesgue measure is beyond the scope of this book.

2.2.4 Measure zero sets

Understanding the measure perspective on probability allows us to understand another important concept of probability, namely **measure zero sets**. To introduce this concept, we pose the question: What is the probability of obtaining a single point, say $\{0.5\}$, when the sample space is $\Omega = [0, 1]$?

The answer to this question is rooted in the **compatibility** between the measure and the sample space. In other words, the measure has to be meaningful for the events in the sample space. Using $\Omega = [0, 1]$, since Ω is an interval, an appropriate measure would be the length of this interval. You may add different weighting functions to define your measure, but ultimately, the measure must be an integral. If you use a "counter" as a measure, then the counter and the interval are not compatible because you cannot count on the real line.

Now, suppose that we define a measure for $\Omega = [0, 1]$ using a weighting function $f(x)$. This measure is determined by an integration. Then, for $E = \{0.5\}$, the measure is

$$\mathbb{P}[E] = \mathbb{P}[\{0.5\}] = \int_{0.5}^{0.5} f(x)\, dx = 0.$$

In fact, for any weighting function the integral will be zero because the length of the set E is zero.[1] An event that gives us zero probability is known as an **event with measure 0**. **Figure 2.20** shows an example.

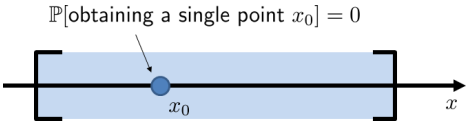


Figure 2.20: The probability of obtaining a single point in a continuous interval is zero.

[1] We assume that f is continuous throughout $[0, 1]$. If f is discontinuous at $x = 0.5$, some additional considerations will apply.

What are measure zero sets?

- A set E (non-empty) is called a measure zero set when $\mathbb{P}[E] = 0$.
- For example, $\{0\}$ is a measure zero set when we use a continuous measure $\mathbb{F}$.
- But $\{0\}$ can have a positive measure when we use a discrete measure $\mathbb{G}$.

Example 2.36(a). Consider a fair die with $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. Then the set $\{\boxdot\}$ has a probability of $\frac{1}{6}$. The sample space does not have a measure zero event because the measure we use is a counter.

Example 2.36(b). Consider an interval with $\Omega = [1, 6]$. Then the set $\{1\}$ has measure 0 because it is an isolated point with respect to the sample space.

Example 2.36(c). For any intervals, $\mathbb{P}[[a, b]] = \mathbb{P}[(a, b)]$ because the two end points have measure zero: $\mathbb{P}[\{a\}] = \mathbb{P}[\{b\}] = 0$.

Formal definitions of measure zero sets

The following discussion of the formal definitions of measure zero sets is optional for the first reading of this book.

We can formally define measure zero sets as follows:

Definition 2.19. *Let Ω be the sample space. A set $A \in \Omega$ is said to have* **measure zero** *if for any given $\epsilon > 0$,*

- *There exists a countable number of subsets A_n such that $A \subseteq \cup_{n=1}^{\infty} A_n$, and*
- $\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \epsilon$.

You may need to read this definition carefully. Suppose we have an event A. We construct a set of neighbors $A_1, \ldots, A_\infty$ such that A is included in the union $\cup_{n=1}^{\infty} A_n$. If the sum of the all $\mathbb{P}[A_n]$ is still less than ϵ, then the set A will have a measure zero.

To understand the difference between a measure for a continuous set and a countable set, consider **Figure 2.21**. On the left side of **Figure 2.21** we show an interval Ω in which there is an isolated point x_0. The measure for this Ω is the length of the interval (relative to whatever weighting function you use). We define a small neighborhood $A_0 = (x_0 - \frac{\epsilon}{2},\ x_0 + \frac{\epsilon}{2})$ surrounding x_0. The length of this interval is not more than ϵ. We then shrink ϵ. However, regardless of how small ϵ is, since x_0 is an isolated point, it is always included in the neighborhood. Therefore, the definition is satisfied, and so $\{x_0\}$ has measure zero.

Example 2.37. Let $\Omega = [0, 1]$. The set $\{0.5\} \subset \Omega$ has measure zero, i.e., $\mathbb{P}[\{0.5\}] = 0$. To see this, we draw a small interval around 0.5, say $[0.5 - \epsilon/3, 0.5 + \epsilon/3]$. Inside this interval, there is really nothing to measure besides the point 0.5. Thus we have found an interval such that it contains 0.5, and the probability is $\mathbb{P}[[0.5 - \epsilon/3, 0.5 + \epsilon/3]] =$

$2\epsilon/3 < \epsilon$. Therefore, by definition, the set $\{0.5\}$ has measure 0.

The situation is very different for the right-hand side of **Figure 2.21**. Here, the measure is not the length but a counter. So if we create a neighborhood surrounding the isolated point x_0, we can always make a count. As a result, if you shrink ϵ to become a very small number (in this case less than $\frac{1}{4}$), then $\mathbb{P}[\{x_0\}] < \epsilon$ will no longer be true. Therefore, the set $\{x_0\}$ has a non-zero measure when we use the counter as the measure.

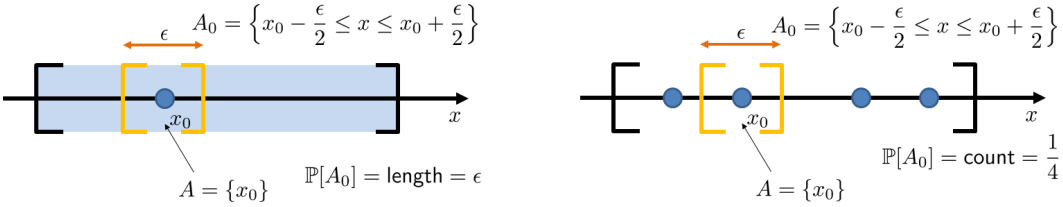


Figure 2.21: [Left] For a continuous sample space, a single point event $\{x_0\}$ can always be surrounded by a neighborhood A_0 whose size $\mathbb{P}[A_0] < \epsilon$. [Right] If you change the sample space to discrete elements, then a single point event $\{x_0\}$ can still be surrounded by a neighborhood A_0. However, the size $\mathbb{P}[A_0] = 1/4$ is a fixed number and will not work for *any* ϵ.

When we make probabilistic claims without considering the measure zero sets, we say that an event happens **almost surely**.

Definition 2.20. *An event $A \in \mathbb{R}$ is said to hold* **almost surely (a.s.)** *if*

$$\mathbb{P}[A] = 1 \tag{2.21}$$

except for all measure zero sets in $\mathbb{R}$.

Therefore, if a set A contains measure zero subsets, we can simply ignore them because they do not affect the probability of events. In this book, we will omit "a.s." if the context is clear.

Example 2.38(a). Let $\Omega = [0, 1]$. Then $\mathbb{P}[(0, 1)] = 1$ almost surely because the points 0 and 1 have measure zero in Ω.

Example 2.38(b). Let $\Omega = \{x \mid x^2 \le 1\}$ and let $A = \{x \mid x^2 < 1\}$. Then $\mathbb{P}[A] = 1$ almost surely because the circumference has measure zero in Ω.

Practice Exercise 2.12. Let $\Omega = \{f : \mathbb{R} \to [-1, 1] \,|\, f(t) = \cos(\omega_0 t + \theta)\}$, where ω_0 is a fixed constant and θ is random. Construct a measure zero event and an almost sure event.

Solution. Let

$$E = \{f : \mathbb{R} \to [-1, 1] \,|\, f(t) = \cos(\omega_0 t + k\pi/2)\}$$

for any integer k. That is, E contains all the functions with a phase of $\pi/2$, $2\pi/2$, $3\pi/2$, etc. Then E will have measure zero because it is a countable set of isolated functions. The event E^c will have probability $\mathbb{P}[E^c] = 1$ almost surely because E has measure

zero.

This is the end of the discussion. Please join us again.

2.2.5 Summary of the probability space

After the preceding long journey through theory, let us summarize.

First, it is extremely important to understand our slogan: **probability is a measure of the size of a set**. This slogan is precise, but it needs clarification. When we say probability is a **measure**, we are thinking of it as being the probability law $\mathbb{P}$. Of course, in practice, we always think of probability as the **number** returned by the measure. However, the difference is not crucial. Also, "size" not only means the number of elements in the set, but it also means the relative weight of the set in the sample space. For example, if we use a weight function to weigh the set elements, then size would refer to the overall weight of the set.

When we put all these pieces together, we can understand why a probability space must consist of the three components

$$(\Omega, \mathcal{F}, \mathbb{P}), \tag{2.22}$$

where Ω is the sample space that defines all possible outcomes, $\mathcal{F}$ is the event space generated from Ω, and $\mathbb{P}$ is the probability law that maps an event to a number in $[0, 1]$. Can we drop one or more of the three components? We cannot! If we do not specify the sample space Ω, then there is no way to define the events. If we do not have a complete event space $\mathcal{F}$, then some events will become undefined, and further, if the probability law is applied only to outcomes, we will not be able to define the probability for events. Finally, if we do not specify the probability law, then we do not have a way to assign probabilities.

2.3 Axioms of Probability

We now turn to a deeper examination of the properties. Our motivation is simple. While the definition of probability law has achieved its goal of assigning a probability to an event, there must be restrictions on how the assignment can be made. For example, if we set $\mathbb{P}[\{H\}] = 1/3$, then $\mathbb{P}[\{T\}]$ must be $2/3$; otherwise, the sum of having a head and a tail will be greater than 1. The necessary restrictions on assigning a probability to an event are collectively known as the **axioms of probability**.

Definition 2.21. *A* **probability law** *is a function* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *that maps an event* A *to a real number in* $[0, 1]$*. The function must satisfy the* **axioms of probability***:*

I. **Non-negativity***:* $\mathbb{P}[A] \geq 0$*, for any* $A \subseteq \Omega$*.*

II. **Normalization***:* $\mathbb{P}[\Omega] = 1$*.*

III. **Additivity***: For any disjoint sets* $\{A_1, A_2, \ldots\}$*, it must be true that*

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]. \tag{2.23}$$

An axiom is a proposition that serves as a premise or starting point in a logical system. Axioms are not definitions, nor are they theorems. They are believed to be true or true within a certain context. In our case, the axioms are true within the context of Bayesian probability. The Kolmogorov probability relies on another set of axioms. We will not dive into the details of these historical issues; in this book, we will confine our discussion to the three axioms given above.

2.3.1 Why these three probability axioms?

Why do we need three axioms? Why not just two axioms? Why these three particular axioms? The reasons are summarized in the box below.

Why these three axioms?

- Axiom I (Non-negativity) ensures that probability is never negative.
- Axiom II (Normalization) ensures that probability is never greater than 1.
- Axiom III (Additivity) allows us to add probabilities when two events do not overlap.

Axiom I is called the **non-negativity** axiom. It ensures that a probability value cannot be negative. Non-negativity is a must for probability. It is meaningless to say that the probability of getting an event is a negative number.

Axiom II is called the **normalization** axiom. It ensures that the probability of observing all possible outcomes is 1. This gives the upper limit of the probability. The upper limit does not have to be 1. It could be 10 or 100. As long as we are consistent about this upper limit, we are good. However, for historical reasons and convenience, we choose 1 to be the upper limit.

Axiom III is called the **additivity** axiom and is the most critical one among the three. The additivity axiom defines how set operations can be translated into probability operations. In a nutshell, it says that if we have a set of disjoint events, the probabilities can be added. From the measure perspective, Axiom III makes sense because if $\mathbb{P}$ measures the size of an event, then two disjoint events should have their probabilities added. If two disjoint events do not allow their probabilities to be added, then there is no way to measure a combined event. Similarly, if the probabilities can somehow be added even for overlapping events, there will be inconsistencies because there is no systematic way to handle the overlapping regions.

The **countable additivity** stated in Axiom III can be applied to both a finite number or an infinite number of sets. The finite case states that for any two disjoint sets A and B, we have

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]. \tag{2.24}$$

In other words, if A and B are disjoint, then the probability of observing either A or B is the sum of the two individual probabilities. **Figure 2.22** illustrates this idea.

Example 2.39. Let's see why Axiom III is critical. Consider throwing a fair die with $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. The probability of getting $\{\boxdot, \boxdot\}$ is

$$\mathbb{P}[\{\boxdot, \boxdot\}] = \mathbb{P}[\{\boxdot\} \cup \{\boxdot\}] = \mathbb{P}[\{\boxdot\}] + \mathbb{P}[\{\boxdot\}] = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}.$$

In this equation, the second equality holds because the events $\{\boxdot\}$ and $\{\boxdot\}$ are disjoint. If we do not have Axiom III, then we cannot **add** probabilities.

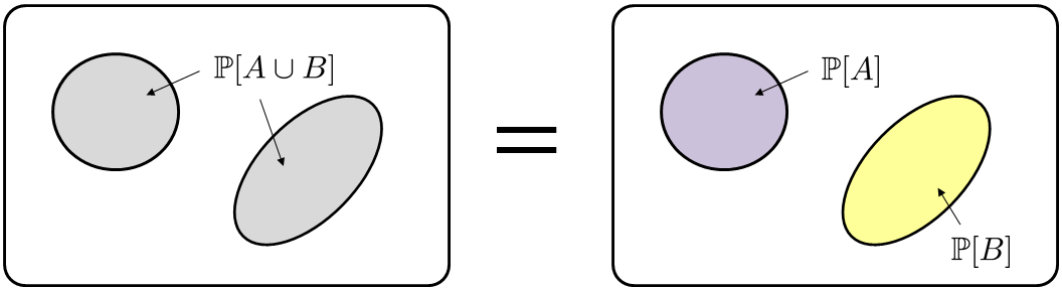


Figure 2.22: Axiom III says $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$ if $A \cap B = \emptyset$.

2.3.2 Axioms through the lens of measure

Axioms are "rules" we must abide by when we construct a measure. Therefore, any valid measure must be compatible with the axioms, regardless of whether we have a weighting function or not. In the following two examples, we will see how the weighting functions are used in the axioms.

Example 2.40. Consider a sample space with $\Omega = \{\clubsuit, \heartsuit, \maltese\}$. The probability for each outcome is

$$\mathbb{P}[\{\clubsuit\}] = \frac{2}{6}, \quad \mathbb{P}[\{\heartsuit\}] = \frac{1}{6}, \quad \mathbb{P}[\{\maltese\}] = \frac{3}{6}.$$

Suppose we construct two disjoint events $E_1 = \{\clubsuit, \heartsuit\}$ and $E_2 = \{\maltese\}$. Then Axiom III says

$$\mathbb{P}[E_1 \cup E_2] = \mathbb{P}[E_1] + \mathbb{P}[E_2] = \left(\frac{2}{6} + \frac{1}{6}\right) + \frac{3}{6} = 1.$$

Note that in this calculation, the measure $\mathbb{P}$ is still a measure $\mathbb{P}$. If we endow it with a nonuniform weight function, then $\mathbb{P}$ applies the corresponding weights to the corresponding outcomes. This process is compatible with the axioms. See **Figure 2.23** for a pictorial illustration.

Example 2.41. Suppose the sample space is an interval $\Omega = [0, 1]$. The two events are $E_1 = [a, b]$ and $E_2 = [c, d]$. Assume that the measure $\mathbb{P}$ uses a weighting function $f(x)$. Then, by Axiom III, we know that

$$\begin{aligned}\mathbb{P}[E_1 \cup E_2] &= \mathbb{P}[E_1] + \mathbb{P}[E_2] \\ &= \mathbb{P}[[a, b]] + \mathbb{P}[[c, d]] \qquad \text{(by Axiom 3)} \\ &= \int_a^b f(x)\, dx + \int_c^d f(x)\, dx, \qquad \text{(apply the measure).}\end{aligned}$$

As you can see, there is no conflict between the axioms and the measure. **Figure 2.24** illustrates this example.

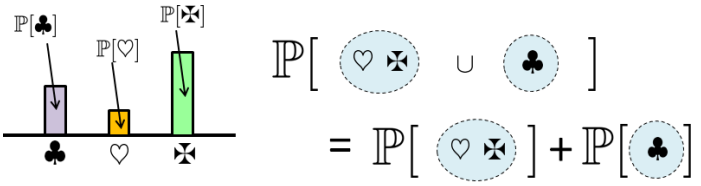


Figure 2.23: Applying weighting functions to the measures: Suppose we have three elements in the set. To compute the probability $\mathbb{P}[\{\heartsuit, ✠\} \cup \{\clubsuit\}]$, we can write it as the sum of $\mathbb{P}[\{\heartsuit, ✠\}]$ and $\mathbb{P}[\{\clubsuit\}]$.

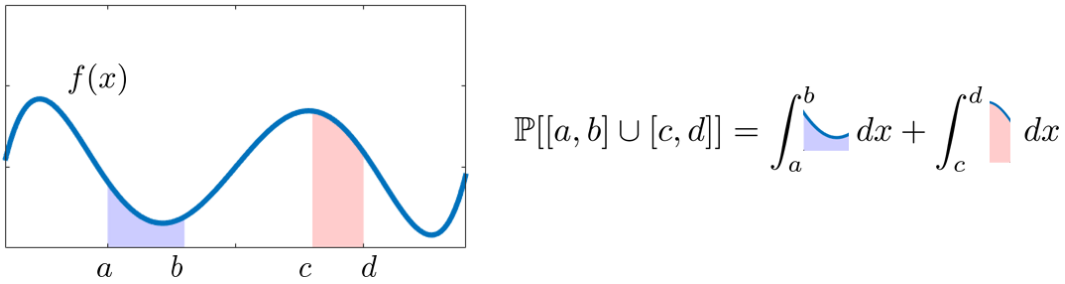


Figure 2.24: The axioms are compatible with the measure, even if we use a weighting function.

2.3.3 Corollaries derived from the axioms

The union of A and B is equivalent to the logical operator "OR". Once the logical operation "OR" is defined, all other logical operations can be defined. The following corollaries are examples.

Corollary 2.1. *Let $A \in \mathcal{F}$ be an event. Then,*

(a) $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$.

(b) $\mathbb{P}[A] \leq 1$.

(c) $\mathbb{P}[\emptyset] = 0$.

Proof. (a) Since $\Omega = A \cup A^c$, by finite additivity we have $\mathbb{P}[\Omega] = \mathbb{P}[A \cup A^c] = \mathbb{P}[A] + \mathbb{P}[A^c]$. By the normalization axiom, we have $\mathbb{P}[\Omega] = 1$. Therefore, $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$.

(b) We prove by contradiction. Assume $\mathbb{P}[A] > 1$. Consider the complement A^c where $A \cup A^c = \Omega$. Since $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$, we must have $\mathbb{P}[A^c] < 0$ because by hypothesis $\mathbb{P}[A] > 1$. But $\mathbb{P}[A^c] < 0$ violates the non-negativity axiom. So we must have $\mathbb{P}[A] \leq 1$.

(c) Since $\Omega = \Omega \cup \emptyset$, by the first corollary we have $\mathbb{P}[\emptyset] = 1 - \mathbb{P}[\Omega] = 0$.

□

Corollary 2.2 (**Unions of Two Non-Disjoint Sets**). *For any A and B in $\mathcal{F}$,*

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]. \tag{2.25}$$

This statement is different from Axiom III because A and B are not necessarily disjoint.

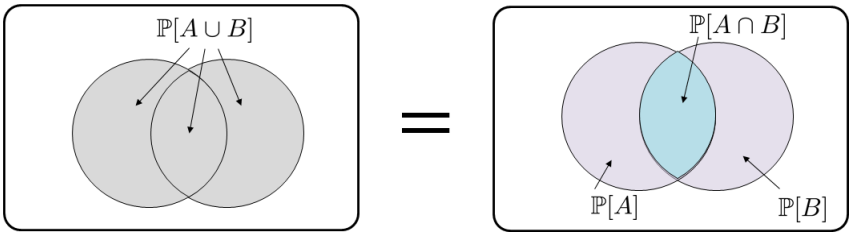


Figure 2.25: For any A and B, $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$.

Proof. First, observe that $A \cup B$ can be partitioned into three disjoint subsets as $A \cup B = (A \backslash B) \cup (A \cap B) \cup (B \backslash A)$. Since $A \backslash B = A \cap B^c$ and $B \backslash A = B \cap A^c$, by finite additivity we have that

$$\begin{aligned}
\mathbb{P}[A \cup B] &= \mathbb{P}[A \backslash B] + \mathbb{P}[A \cap B] + \mathbb{P}[B \backslash A] = \mathbb{P}[A \cap B^c] + \mathbb{P}[A \cap B] + \mathbb{P}[B \cap A^c] \\
&\overset{(a)}{=} \mathbb{P}[A \cap B^c] + \mathbb{P}[A \cap B] + \mathbb{P}[B \cap A^c] + \mathbb{P}[A \cap B] - \mathbb{P}[A \cap B] \\
&\overset{(b)}{=} \mathbb{P}[A \cap (B^c \cup B)] + \mathbb{P}[(A^c \cup A) \cap B] - \mathbb{P}[A \cap B] \\
&= \mathbb{P}[A \cap \Omega] + \mathbb{P}[\Omega \cap B] - \mathbb{P}[A \cap B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B],
\end{aligned}$$

where in (a) we added and subtracted a term $\mathbb{P}[A \cap B]$, and in (b) we used finite additivity so that $\mathbb{P}[A \cap B^c] + \mathbb{P}[A \cap B] = \mathbb{P}[(A \cap B^c) \cup (A \cap B)] = \mathbb{P}[A \cap (B^c \cup B)]$.

□

Example 2.42. The corollary is easy to understand if we consider the following example. Let $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$ be the sample space of a fair die. Let $A = \{\boxdot, \boxdot, \boxdot\}$ and $B = \{\boxdot, \boxdot, \boxdot\}$. Then

$$\mathbb{P}[A \cup B] = \mathbb{P}[\{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}] = \frac{5}{6}.$$

We can also use the corollary to obtain the same result:

$$\begin{aligned}\mathbb{P}[A \cup B] &= \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \\ &= \mathbb{P}[\{\boxdot, \boxdot, \boxdot\}] + \mathbb{P}[\{\boxdot, \boxdot, \boxdot\}] - \mathbb{P}[\{\boxdot\}] \\ &= \frac{3}{6} + \frac{3}{6} - \frac{1}{6} = \frac{5}{6}.\end{aligned}$$

Corollary 2.3 (Inequalities)**.** *Let A and B be two events in $\mathcal{F}$. Then,*

(a) $\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$. *(Union Bound)*

(b) If $A \subseteq B$, *then* $\mathbb{P}[A] \leq \mathbb{P}[B]$.

Proof. (a) Since $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$ and by non-negativity axiom $\mathbb{P}[A \cap B] \geq 0$, we must have $\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$. (b) If $A \subseteq B$, then there exists a set $B \backslash A$ such that $B = A \cup (B \backslash A)$. Therefore, by finite additivity we have $\mathbb{P}[B] = \mathbb{P}[A] + \mathbb{P}[B \backslash A] \geq \mathbb{P}[A]$. Since $\mathbb{P}[B \backslash A] \geq 0$, it follows that $\mathbb{P}[A] + \mathbb{P}[B \backslash A] \geq \mathbb{P}[A]$. Thus we have $\mathbb{P}[B] \geq \mathbb{P}[A]$.

□

Union bound is a frequently used tool for analyzing probabilities when the intersection $A \cap B$ is difficult to evaluate. Part (b) is useful when considering two events of different "sizes." For example, in the bus-waiting example, if we let $A = \{t \leq 5\}$, and $B = \{t \leq 10\}$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$ because we have to wait for the first 5 minutes to go into the remaining 5 minutes.

Practice Exercise 2.13. Let the events A and B have $\mathbb{P}[A] = x$, $\mathbb{P}[B] = y$ and $\mathbb{P}[A \cup B] = z$. Find the following probabilities: $\mathbb{P}[A \cap B]$, $\mathbb{P}[A^c \cup B^c]$, and $\mathbb{P}[A \cap B^c]$.

Solution.

(a) Note that $z = \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$. Thus, $\mathbb{P}[A \cap B] = x + y - z$.

(b) We can take the complement to obtain the result:

$$\mathbb{P}[A^c \cup B^c] = 1 - \mathbb{P}[(A^c \cup B^c)^c] = 1 - \mathbb{P}[A \cap B] = 1 - x - y + z.$$

(c) $\mathbb{P}[A \cap B^c] = \mathbb{P}[A] - \mathbb{P}[A \cap B] = x - (x + y - z) = z - y$.

Practice Exercise 2.14. Consider a sample space

$$\Omega = \{f : \mathbb{R} \to \mathbb{R} \,|\, f(x) = ax, \text{for all } a \in \mathbb{R}, x \in \mathbb{R}\}.$$

There are two events: $A = \{f \,|\, f(x) = ax,\ a \geq 0\}$, and $B = \{f \,|\, f(x) = ax,\ a \leq 0\}$. So, basically, A is the set of all straight lines with positive slope, and B is the set of straight lines with negative slope. Show that the union bound is tight.

Solution. First of all, we note that

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B].$$

The intersection is

$$\mathbb{P}[A \cap B] = \mathbb{P}[\{f \mid f(x) = 0\}].$$

Since this is a point set in the real line, it has measure zero. Thus, $\mathbb{P}[A \cap B] = 0$ and hence $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$. So the union bound is tight.

Closing remark. The development of today's probability theory is generally credited to Andrey Kolmogorov's 1933 book *Foundations of the Theory of Probability*. We close this section by citing one of the tables of the book. The table summarizes the correspondence between set theory and random events.

Theory of sets	Random events
A and B are disjoint, i.e., $A \cap B = \emptyset$	Events A and B are incompatible
$A_1 \cap A_2 \cdots \cap A_N = \emptyset$	Events $A_1, \ldots, A_N$ are incompatible
$A_1 \cap A_2 \cdots \cap A_N = X$	Event X is defined as the simultaneous occurrence of events $A_1, \ldots, A_N$
$A_1 \cup A_2 \cdots \cup A_N = X$	Event X is defined as the occurrence of at least one of the events $A_1, \ldots, A_N$
A^c	The opposite event A^c consisting of the non-occurrence of event A
$A = \emptyset$	Event A is impossible
$A = \Omega$	Event A must occur
$A_1, \ldots, A_N$ form a partition of Ω	The experiment consists of determining which of the events $A_1, \ldots, A_N$ occurs
$B \subset A$	From the occurrence of event B follows the inevitable occurrence of A

Table 2.2: Kolmogorov's summary of set theory results and random events.

2.4 Conditional Probability

In many practical data science problems, we are interested in the relationship between two or more events. For example, an event A may cause B to happen, and B may cause C to happen. A legitimate question in probability is then: If A has happened, what is the probability that B also happens? Of course, if A and B are correlated events, then knowing one event can tell us something about the other event. If the two events have no relationship, knowing one event will not tell us anything about the other.

In this section, we study the concept of **conditional probability**. There are three subtopics in this section. We summarize the key points below.

The three main messages of this section are:

- Section 2.4.1: **Conditional probability**. Conditional probability of A given B is $\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$.
- Section 2.4.2: **Independence**. Two events are **independent** if the occurrence of one does not influence the occurrence of the other: $\mathbb{P}[A|B] = \mathbb{P}[A]$.
- Section 2.4.3: **Bayes' theorem and the law of total probability**. Bayes' theorem allows us to switch the order of the conditioning: $\mathbb{P}[A|B]$ vs. $\mathbb{P}[B|A]$, whereas the law of total probability allows us to decompose an event into smaller events.

2.4.1 Definition of conditional probability

We start by defining **conditional probability**.

Definition 2.22. *Consider two events A and B. Assume $\mathbb{P}[B] \neq 0$. The* **conditional probability** *of A given B is*

$$\mathbb{P}[A \mid B] \stackrel{\text{def}}{=} \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}. \tag{2.26}$$

According to this definition, the conditional probability of A given B is the ratio of $\mathbb{P}[A \cap B]$ to $\mathbb{P}[B]$. It is the probability that A happens when we know that B has already happened. Since B has already happened, the event that A has also happened is represented by $A \cap B$. However, since we are only interested in the relative probability of A with respect to B, we need to normalize using B. This can be seen by comparing $\mathbb{P}[A \mid B]$ and $\mathbb{P}[A \cap B]$:

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \quad \text{and} \quad \mathbb{P}[A \cap B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[\Omega]}. \tag{2.27}$$

The difference is illustrated in **Figure 2.26**: The intersection $\mathbb{P}[A \cap B]$ calculates the overlapping area of the two events. We make no assumptions about the cause-effect relationship.

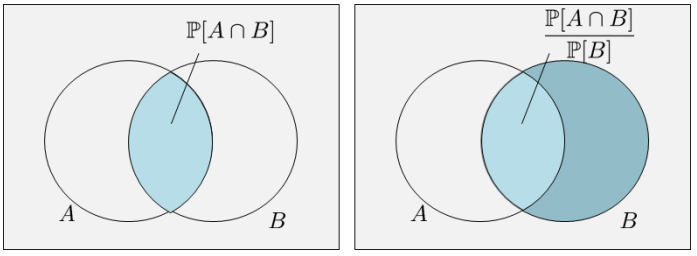


Figure 2.26: Illustration of conditional probability and its comparison with $\mathbb{P}[A \cap B]$.

What justifies this ratio? Suppose that B has already happened. Then, anything outside B will immediately become irrelevant as far as the relationship between A and B is concerned. So when we ask: "What is the probability that A happens given that B has happened?", we are effectively asking for the probability that $A \cap B$ happens under the

condition that B has happened. Note that we need to consider $A \cap B$ because we know that B has already happened. If we take A only, then there exists a region $A \backslash B$ which does not contain anything about B. However, since we know that B has happened, $A \backslash B$ is impossible. In other words, among the elements of A, only those that appear in $A \cap B$ are meaningful.

Example 2.43. Let

$$A = \{\text{Purdue gets Big Ten championship}\},$$
$$B = \{\text{Purdue wins 15 games consecutively}\}.$$

In this example,

$$\begin{aligned}
\mathbb{P}[A] &= \text{Prob. that Purdue gets the championship,} \\
\mathbb{P}[B] &= \text{Prob. that Purdue wins 15 games consecutively,} \\
\mathbb{P}[A \cap B] &= \text{Prob. that Purdue gets the championship and wins 15 games,} \\
\mathbb{P}[A \,|\, B] &= \text{Prob. that Purdue gets the championship given that} \\
&\qquad \text{Purdue won 15 games.}
\end{aligned}$$

If Purdue has won 15 games consecutively, then it is unlikely that Purdue will get the championship because the sample space of all possible competition results is large. However, if we have already won 15 games consecutively, then the denominator of the probability becomes much smaller. In this case, the conditional probability is high.

Example 2.44. Consider throwing a die. Let

$$A = \{\text{getting a 3}\} \quad \text{and} \quad B = \{\text{getting an odd number}\}.$$

Find $\mathbb{P}[A \,|\, B]$ and $\mathbb{P}[B \,|\, A]$.

Solution. The following probabilities are easy to calculate:

$$\mathbb{P}[A] = \mathbb{P}[\{\boxdot\}] = \frac{1}{6}, \qquad \text{and} \qquad \mathbb{P}[B] = \mathbb{P}[\{\boxdot, \boxdot, \boxdot\}] = \frac{3}{6}.$$

Also, the intersection is

$$\mathbb{P}[A \cap B] = \mathbb{P}[\{\boxdot\}] = \frac{1}{6}.$$

Given these values, the conditional probability of A given B can be calculated as

$$\mathbb{P}[A \,|\, B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$$

In other words, if we know that we have an odd number, then the probability of obtaining a 3 has to be computed over $\{\boxdot, \boxdot, \boxdot\}$, which give us a probability $\frac{1}{3}$. If we

do not know that we have an odd number, then the probability of obtaining a 3 has to be computed from the sample space $\{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$, which will give us $\frac{1}{6}$.

The other conditional probability is

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = 1.$$

Therefore, if we know that we have rolled a 3, then the probability for this number being an odd number is 1.

Example 2.45. Consider the situation shown in **Figure 2.27**. There are 12 points with equal probabilities of happening. Find the probabilities $\mathbb{P}[A|B]$ and $\mathbb{P}[B|A]$.

Solution. In this example, we can first calculate the individual probabilities:

$$\mathbb{P}[A] = \frac{5}{12}, \quad \text{and} \quad \mathbb{P}[B] = \frac{6}{12}, \quad \text{and} \quad \mathbb{P}[A \cap B] = \frac{2}{12}.$$

Then the conditional probabilities are

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\frac{2}{12}}{\frac{6}{12}} = \frac{1}{3},$$
$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = \frac{\frac{2}{12}}{\frac{5}{12}} = \frac{2}{5}.$$

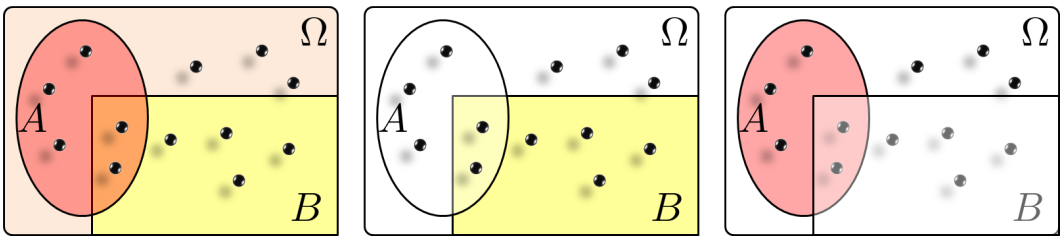


Figure 2.27: Visualization of Example 2.45: [Left] All the sets. [Middle] $P(A|B)$ is the ratio between dots inside the light yellow region over those in yellow, which is $\frac{2}{6}$. [Right] $\mathbb{P}[A|B]$ is the ratio between dots inside the light pink region over those in pink, which is $\frac{2}{5}$.

Example 2.46. Consider a tetrahedral (4-sided) die. Let X be the first roll and Y be the second roll. Let B be the event that $\min(X, Y) = 2$ and M be the event that $\max(X, Y) = 3$. Find $\mathbb{P}[M|B]$.

Solution. As shown in **Figure 2.28**, the event B is highlighted in green. (Why?) Similarly, the event M is highlighted in blue. (Again, why?) Therefore, the probability

is

$$\mathbb{P}[M|B] = \frac{\mathbb{P}[M \cap B]}{\mathbb{P}[B]} = \frac{\frac{2}{16}}{\frac{5}{16}} = \frac{2}{5}.$$

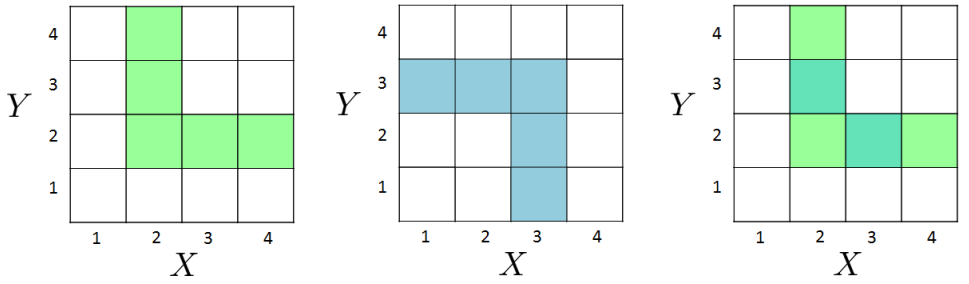


Figure 2.28: Visualization of Example 2.46. [Left] Event B. [Middle] Event M. [Right] $\mathbb{P}(M|B)$ is the ratio of the number of blue squares inside the green region to the total number of green squares, which is $\frac{2}{5}$.

Remark. Notice that if $\mathbb{P}[B] \leq \mathbb{P}[\Omega]$, then $\mathbb{P}[A \mid B]$ is always larger than or equal to $\mathbb{P}[A \cap B]$, i.e.,

$$\mathbb{P}[A|B] \geq \mathbb{P}[A \cap B].$$

Conditional probabilities are legitimate probabilities

Conditional probabilities are legitimate probabilities. That is, given B, the probability $\mathbb{P}[A|B]$ satisfies Axioms I, II, III.

Theorem 2.6. *Let $\mathbb{P}[B] > 0$. The conditional probability $\mathbb{P}[A \mid B]$ satisfies Axioms I, II, and III.*

Proof. Let's check the axioms:

- Axiom I: We want to show

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \geq 0.$$

  Since $\mathbb{P}[B] > 0$ and Axiom I requires $\mathbb{P}[A \cap B] \geq 0$, we therefore have $\mathbb{P}[A \mid B] \geq 0$.

- Axiom II:

$$\begin{aligned}\mathbb{P}[\Omega \mid B] &= \frac{\mathbb{P}[\Omega \cap B]}{\mathbb{P}[B]} \\ &= \frac{\mathbb{P}[B]}{\mathbb{P}[B]} = 1.\end{aligned}$$

- Axiom III: Consider two disjoint sets A and C. Then,

$$\begin{aligned}\mathbb{P}[A \cup C \,|\, B] &= \frac{\mathbb{P}[(A \cup C) \cap B]}{\mathbb{P}[B]} \\ &= \frac{\mathbb{P}[(A \cap B) \cup (C \cap B)]}{\mathbb{P}[B]} \\ &\overset{(a)}{=} \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} + \frac{\mathbb{P}[C \cap B]}{\mathbb{P}[B]} \\ &= \mathbb{P}[A|B] + \mathbb{P}[C|B],\end{aligned}$$

where (a) holds because if A and C are disjoint then $(A \cap B) \cap (C \cap B) = \emptyset$.

□

To summarize this subsection, we highlight the essence of conditional probability.

What are conditional probabilities?

- Conditional probability of A given B is the ratio $\frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$.
- It is again a **measure**. It measures the relative size of A **inside** B.
- Because it is a measure, it must satisfy the three axioms.

2.4.2 Independence

Conditional probability deals with situations where two events A and B are related. What if the two events are unrelated? In probability, we have a technical term for this situation: statistical **independence**.

Definition 2.23. *Two events A and B are statistically* **independent** *if*

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]. \tag{2.28}$$

Why define independence in this way? Recall that $\mathbb{P}[A \,|\, B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$. If A and B are independent, then $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$ and so

$$\mathbb{P}[A \,|\, B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[A]\,\mathbb{P}[B]}{\mathbb{P}[B]} = \mathbb{P}[A]. \tag{2.29}$$

This suggests an interpretation of independence: If the occurrence of B provides no additional information about the occurrence of A, then A and B are independent.

Therefore, we can define independence via conditional probability:

Definition 2.24. *Let A and B be two events such that $\mathbb{P}[A] > 0$ and $\mathbb{P}[B] > 0$. Then*

A and B are **independent** *if*

$$\mathbb{P}[A \,|\, B] = \mathbb{P}[A] \quad \textit{or} \quad \mathbb{P}[B \,|\, A] = \mathbb{P}[B]. \tag{2.30}$$

The two statements are equivalent as long as $\mathbb{P}[A] > 0$ and $\mathbb{P}[B] > 0$. This is because $\mathbb{P}[A|B] = \mathbb{P}[A \cap B]/\mathbb{P}[B]$. If $\mathbb{P}[A|B] = \mathbb{P}[A]$ then $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$, which implies that $\mathbb{P}[B|A] = \mathbb{P}[A \cap B]/\mathbb{P}[A] = \mathbb{P}[B]$.

A pictorial illustration of independence is given in **Figure 2.29**. The key message is that if two events A and B are independent, then $\mathbb{P}[A|B] = \mathbb{P}[A]$. The conditional probability $\mathbb{P}[A|B]$ is the ratio of $\mathbb{P}[A \cap B]$ over $\mathbb{P}[B]$, which is the intersection over B (the blue set). The probability $\mathbb{P}[A]$ is the yellow set over the sample space Ω.

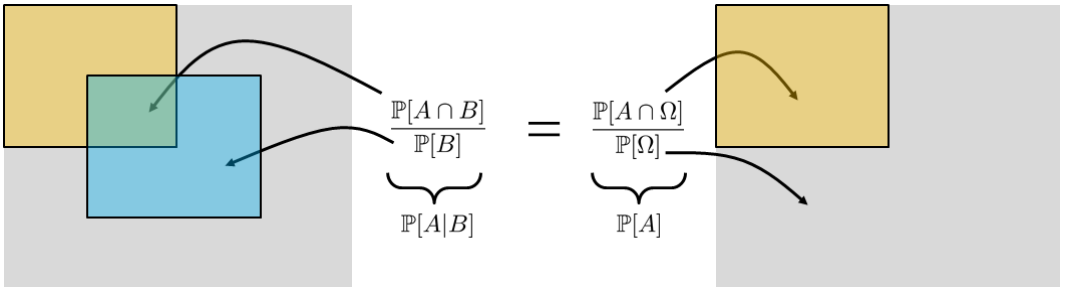


Figure 2.29: Independence means that the conditional probability $\mathbb{P}[A|B]$ is the same as $\mathbb{P}[A]$. This implies that the ratio of $\mathbb{P}[A \cap B]$ over $\mathbb{P}[B]$, and the ratio of $\mathbb{P}[A \cap \Omega]$ over $\mathbb{P}[\Omega]$ are the same.

Disjoint versus independent

$$\text{Disjoint} \not\Leftrightarrow \text{Independent}. \tag{2.31}$$

The statement says that disjoint and independent are two completely different concepts.

If A and B are disjoint, then $A \cap B = \emptyset$. This only implies that $\mathbb{P}[A \cap B] = 0$. However, it says nothing about whether $\mathbb{P}[A \cap B]$ can be factorized into $\mathbb{P}[A]\,\mathbb{P}[B]$. If A and B are independent, then we have $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$. But this does not imply that $\mathbb{P}[A \cap B] = 0$. The only condition under which Disjoint $\Leftrightarrow$ Independence is when $\mathbb{P}[A] = 0$ or $\mathbb{P}[B] = 0$. **Figure 2.30** depicts the situation. When two sets are independent, the conditional probability (which is a ratio) remains unchanged compared to unconditioned probability. When two sets are disjoint, they simply do not overlap.

Practice Exercise 2.15. Throw a die twice. Are A and B independent, where

$$A = \{\text{1st die is 3}\} \quad \text{and} \quad B = \{\text{2nd die is 4}\}.$$

Solution. We can show that

$$\mathbb{P}[A \cap B] = \mathbb{P}[(3,4)] = \tfrac{1}{36}, \qquad \mathbb{P}[A] = \tfrac{1}{6}, \quad \text{and} \quad \mathbb{P}[B] = \tfrac{1}{6}.$$

So $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. Thus, A and B are independent.

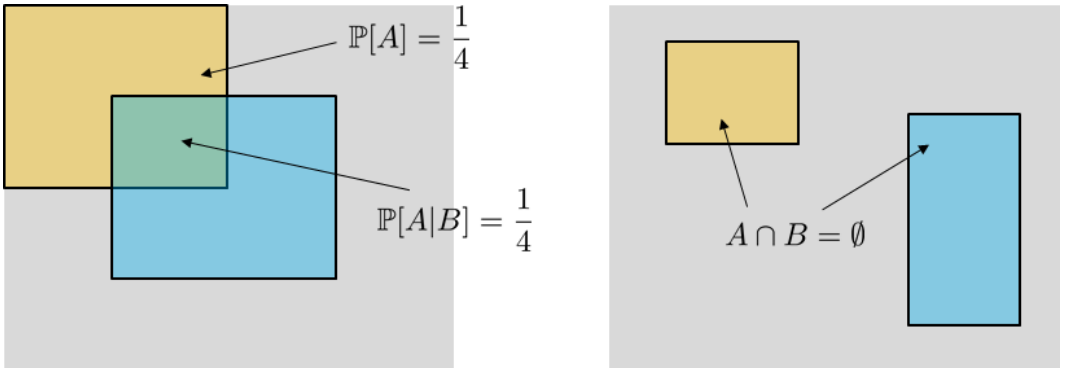


Figure 2.30: Independent means that the conditional probability, which is a ratio, is the same as the unconditioned probability. Disjoint means that the two sets do not overlap.

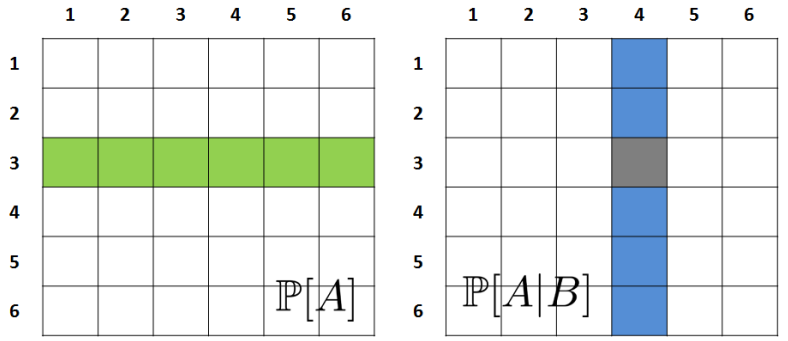


Figure 2.31: The two events A and B are independent because $\mathbb{P}[A] = \frac{1}{6}$ and $\mathbb{P}[A|B] = \frac{1}{6}$.

A pictorial illustration of this example is shown in **Figure 2.31**. The two events are independent because A is one row in the 2D space, which yields a probability of $\frac{1}{6}$. The conditional probability $\mathbb{P}[A|B]$ is the coordinate $(3, 4)$ over the event B, which is a column. It happens that $\mathbb{P}[A|B] = \frac{1}{6}$. Thus, the two events are independent.

Practice Exercise 2.16. Throw a die twice. Are A and B independent?

$$A = \{\text{1st die is 3}\} \quad \text{and} \quad B = \{\text{sum is 7}\}.$$

Solution. Note that

$$\mathbb{P}[A \cap B] = \mathbb{P}[(3,4)] = \tfrac{1}{36}, \qquad \mathbb{P}[A] = \tfrac{1}{6},$$
$$\mathbb{P}[B] = \mathbb{P}[(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)] = \tfrac{1}{6}.$$

So $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$. Thus, A and B are independent.

A pictorial illustration of this example is shown in **Figure 2.32**. Notice that whether the two events intersect is not how we determine independence (that only determines disjoint or

not). The key is whether the conditional probability (which is the ratio) remains unchanged compared to the unconditioned probability.

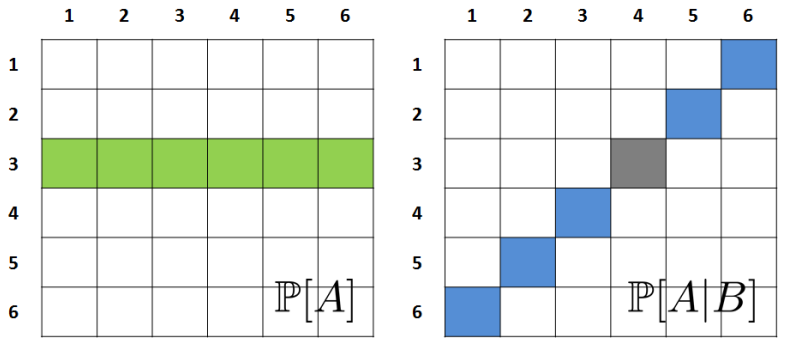


Figure 2.32: The two events A and B are independent because $\mathbb{P}[A] = \frac{1}{6}$ and $\mathbb{P}[A \cap B] = \frac{1}{6}$.

If we let $B = \{\text{sum is } 8\}$, then the situation is different. The intersection $A \cap B$ has a probability $\frac{1}{5}$ relative to B, and therefore $\mathbb{P}[A|B] = \frac{1}{5}$. Hence, the two events A and B are dependent. If you like a more intuitive argument, you can imagine that B has happened, i.e., the sum is 8. Then the probability for the first die to be 1 is 0 because there is no way to construct 8 when the first die is 1. As a result, we have eliminated one choice for the first die, leaving only five options. Therefore, since B has influenced the probability of A, they are dependent.

Practice Exercise 2.17. Throw a die twice. Let

$$A = \{\text{max is } 2\} \quad \text{and} \quad B = \{\text{min is } 2\}.$$

Are A and B independent?

Solution. Let us first list out A and B:

$$A = \{(1,2),(2,1),(2,2)\},$$
$$B = \{(2,2),(2,3),(2,4),(2,5),(2,6),(3,2),(4,2),(5,2),(6,2)\}.$$

Therefore, the probabilities are

$$\mathbb{P}[A] = \frac{3}{36}, \quad \mathbb{P}[B] = \frac{9}{36}, \quad \text{and} \quad \mathbb{P}[A \cap B] = \mathbb{P}[(2,2)] = \frac{1}{36}.$$

Clearly, $\mathbb{P}[A \cap B] \neq \mathbb{P}[A]\mathbb{P}[B]$ and so A and B are dependent.

What is independence?

- Two events are independent when the **ratio** $\mathbb{P}[A \cap B]/\mathbb{P}[B]$ **remains unchanged** compared to $\mathbb{P}[A]$.
- Independence $\neq$ disjoint.

2.4.3 Bayes' theorem and the law of total probability

Theorem 2.7 (**Bayes' theorem**). *For any two events A and B such that $\mathbb{P}[A] > 0$ and $\mathbb{P}[B] > 0$,*

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B \mid A]\,\mathbb{P}[A]}{\mathbb{P}[B]}.$$

Proof. By the definition of conditional probabilities, we have

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \quad \text{and} \quad \mathbb{P}[B \mid A] = \frac{\mathbb{P}[B \cap A]}{\mathbb{P}[A]}.$$

Rearranging the terms yields

$$\mathbb{P}[A \mid B]\mathbb{P}[B] = \mathbb{P}[B \mid A]\mathbb{P}[A],$$

which gives the desired result by dividing both sides by $\mathbb{P}[B]$.

□

Bayes' theorem provides two views of the intersection $\mathbb{P}[A \cap B]$ using two different conditional probabilities. We call $\mathbb{P}[B \mid A]$ the **conditional probability** and $\mathbb{P}[A \mid B]$ the **posterior probability**. The order of A and B is arbitrary. We can also call $\mathbb{P}[A \mid B]$ the conditional probability and $\mathbb{P}[B \mid A]$ the posterior probability. The context of the problem will make this clear.

Bayes' theorem provides a way to switch $\mathbb{P}[A|B]$ and $\mathbb{P}[B|A]$. The next theorem helps us decompose an event into smaller events.

Theorem 2.8 (**Law of Total Probability**). *Let $\{A_1, \ldots, A_n\}$ be a partition of Ω, i.e., $A_1, \ldots, A_n$ are disjoint and $\Omega = A_1 \cup \cdots \cup A_n$. Then, for any $B \subseteq \Omega$,*

$$\mathbb{P}[B] = \sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\ \mathbb{P}[A_i]. \tag{2.32}$$

Proof. We start from the right-hand side.

$$\sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\ \mathbb{P}[A_i] \overset{(a)}{=} \sum_{i=1}^{n} \mathbb{P}[B \cap A_i] \overset{(b)}{=} \mathbb{P}\left[\bigcup_{i=1}^{n} (B \cap A_i)\right]$$
$$\overset{(c)}{=} \mathbb{P}\left[B \cap \left(\bigcup_{i=1}^{n} A_i\right)\right] \overset{(d)}{=} \mathbb{P}[B \cap \Omega] = \mathbb{P}[B],$$

where (a) follows from the definition of conditional probability, (b) is due to Axiom III, (c) holds because of the distributive property of sets, and (d) results from the partition property of $\{A_1, A_2, \ldots, A_n\}$.

□

Interpretation. The law of total probability can be understood as follows. If the sample space Ω consists of disjoint subsets $A_1, \ldots, A_n$, we can compute the probability $\mathbb{P}[B]$ by

summing over its portion $\mathbb{P}[B \cap A_1], \ldots, \mathbb{P}[B \cap A_n]$. However, each intersection can be written as

$$\mathbb{P}[B \cap A_i] = \mathbb{P}[B \mid A_i]\mathbb{P}[A_i]. \tag{2.33}$$

In other words, we write $\mathbb{P}[B \cap A_i]$ as the **conditional probability** $\mathbb{P}[B \mid A_i]$ times the **prior probability** $\mathbb{P}[A_i]$. When we sum all these intersections, we obtain the overall probability. See **Figure 2.33** for a graphical portrayal.

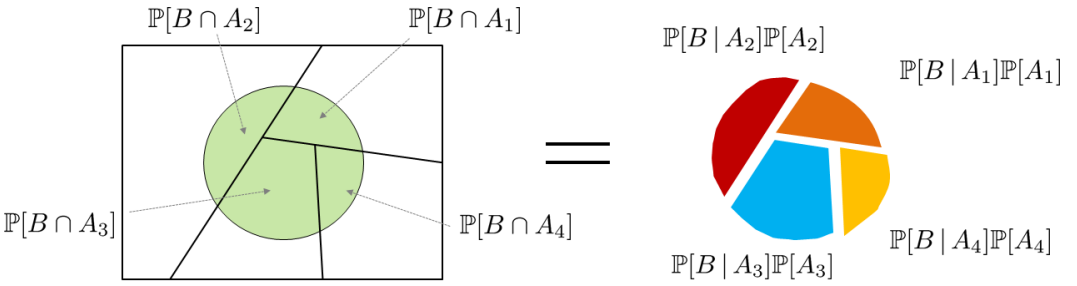


Figure 2.33: The law of total probability decomposes the probability $\mathbb{P}[B]$ into multiple conditional probabilities $\mathbb{P}[B \mid A_i]$. The probability of obtaining each $\mathbb{P}[B \mid A_i]$ is $\mathbb{P}[A_i]$.

Corollary 2.4. *Let $\{A_1, A_2, \ldots, A_n\}$ be a partition of Ω, i.e., $A_1, \ldots, A_n$ are disjoint and $\Omega = A_1 \cup A_2 \cup \cdots \cup A_n$. Then, for any $B \subseteq \Omega$,*

$$\mathbb{P}[A_j \mid B] = \frac{\mathbb{P}[B \mid A_j]\,\mathbb{P}[A_j]}{\sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\,\mathbb{P}[A_i]}. \tag{2.34}$$

Proof. The result follows directly from Bayes' theorem:

$$\mathbb{P}[A_j \mid B] = \frac{\mathbb{P}[B \mid A_j]\,\mathbb{P}[A_j]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B \mid A_j]\,\mathbb{P}[A_j]}{\sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\,\mathbb{P}[A_i]}.$$

□

Example 2.47. Suppose there are three types of players in a tennis tournament: A, B, and C. Fifty percent of the contestants in the tournament are A players, 25% are B players, and 25% are C players. Your chance of beating the contestants depends on the class of the player, as follows:

0.3 against an A player
0.4 against a B player
0.5 against a C player

If you play a match in this tournament, what is the probability of your winning the match? Supposing that you have won a match, what is the probability that you played against an A player?

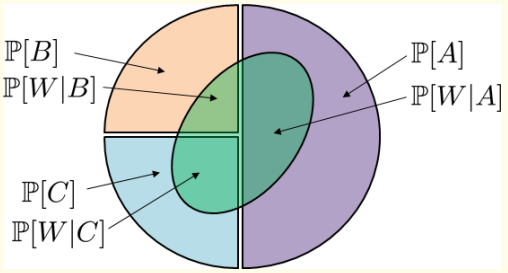
Solution. We first list all the known probabilities. We know from the percentage

of players that

$$\mathbb{P}[A] = 0.5, \quad \mathbb{P}[B] = 0.25, \quad \mathbb{P}[C] = 0.25.$$

Now, let W be the event that you win the match. Then the conditional probabilities are defined as follows:

$$\mathbb{P}[W|A] = 0.3, \quad \mathbb{P}[W|B] = 0.4, \quad \mathbb{P}[W|C] = 0.5.$$



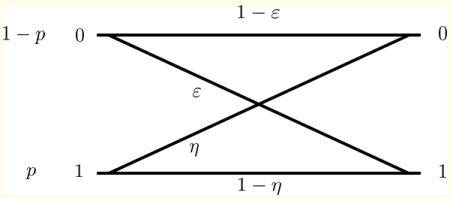
Therefore, by the law of total probability, we can show that the probability of winning the match is

$$\begin{aligned}\mathbb{P}[W] &= \mathbb{P}[W \mid A]\,\mathbb{P}[A] + \mathbb{P}[W \mid B]\,\mathbb{P}[B] + \mathbb{P}[W \mid C]\,\mathbb{P}[C] \\ &= (0.3)(0.5) + (0.4)(0.25) + (0.5)(0.25) = 0.375.\end{aligned}$$

Given that you have won the match, the probability of A given W is

$$\mathbb{P}[A|W] = \frac{\mathbb{P}[W|A]\mathbb{P}[A]}{\mathbb{P}[W]} = \frac{(0.3)(0.5)}{0.375} = 0.4.$$

Example 2.48. Consider the communication channel shown below. The probability of sending a 1 is p and the probability of sending a 0 is $1-p$. Given that 1 is sent, the probability of receiving 1 is $1-\eta$. Given that 0 is sent, the probability of receiving 0 is $1-\varepsilon$. Find the probability that a 1 has been correctly received.



Solution. Define the events

$$\begin{aligned}S_0 &= \text{“0 is sent”}, \quad \text{and} \quad R_0 = \text{“0 is received”}. \\ S_1 &= \text{“1 is sent”}, \quad \text{and} \quad R_1 = \text{“1 is received”}.\end{aligned}$$

Then, the probability that 1 is received is $\mathbb{P}[R_1]$. However, $\mathbb{P}[R_1] \neq 1-\eta$ because $1-\eta$

is the conditional probability that 1 is received given that 1 is sent. It is possible that we receive 1 as a result of an error when 0 is sent. Therefore, we need to consider the probability that both S_0 and S_1 occur. Using the law of total probability we have

$$\begin{aligned}\mathbb{P}[R_1] &= \mathbb{P}[R_1 \mid S_1]\,\mathbb{P}[S_1] + \mathbb{P}[R_1 \mid S_0]\,\mathbb{P}[S_0] \\ &= (1-\eta)p + \varepsilon(1-p).\end{aligned}$$

Now, suppose that we have received 1. What is the probability that 1 was originally sent? This is asking for the posterior probability $\mathbb{P}[S_1 \mid R_1]$, which can be found using Bayes' theorem

$$\mathbb{P}[S_1 \mid R_1] = \frac{\mathbb{P}[R_1 \mid S_1]\,\mathbb{P}[S_1]}{\mathbb{P}[R_1]} = \frac{(1-\eta)p}{(1-\eta)p + \varepsilon(1-p)}.$$

When do we need to use Bayes' theorem and the law of total probability?

- Bayes' theorem **switches** the role of the conditioning, from $\mathbb{P}[A|B]$ to $\mathbb{P}[B|A]$.
  Example:

  $$\mathbb{P}[\text{win the game} \mid \text{play with A}] \quad \text{and} \quad \mathbb{P}[\text{play with A} \mid \text{win the game}].$$

- The law of total probability **decomposes** an event into smaller events.
  Example:

  $$\mathbb{P}[\text{win}] = \mathbb{P}[\text{win} \mid A]\mathbb{P}[A] + \mathbb{P}[\text{win} \mid B]\mathbb{P}[B].$$

2.4.4 The Three Prisoners problem

Now that you are familiar with the concepts of conditional probabilities, we would like to challenge you with the following problem, known as the **Three Prisoners problem**. If you understand how this problem can be resolved, you have mastered conditional probability.

Once upon a time, there were three prisoners A, B, and C. One day, the king decided to pardon two of them and sentence the last one, as in this figure:

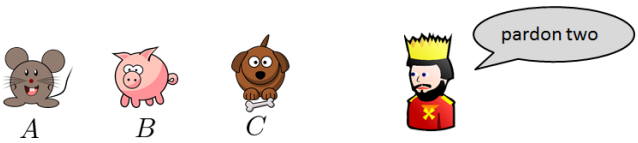


Figure 2.34: The Three Prisoners problem: The king says that he will pardon two prisoners and sentence one.

One of the prisoners, prisoner A, heard the news and wanted to ask a friendly guard about his situation. The guard was honest. He was allowed to tell prisoner A that prisoner B would be pardoned or that prisoner C would be pardoned, but he could not tell A whether he would be pardoned. Prisoner A thought about the problem, and he began to hesitate to ask the guard. Based on his present state of knowledge, his probability of being pardoned

is $\frac{2}{3}$. However, if he asks the guard, this probability will be reduced to $\frac{1}{2}$ because the guard would tell him that one of the two other prisoners would be pardoned, and would tell him which one it would be. Prisoner A reasons that his chance of being pardoned would then drop because there are now only two prisoners left who may be pardoned, as illustrated in **Figure 2.35**:

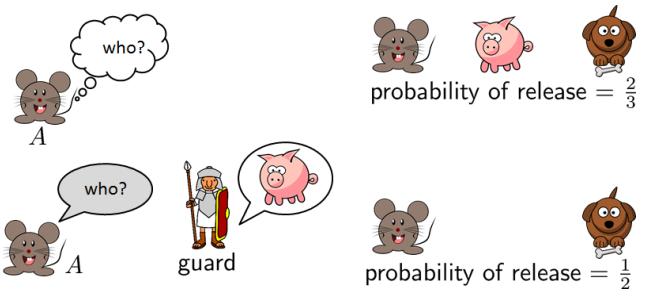


Figure 2.35: The Three Prisoners problem: If you do not ask the guard, your chance of being released is 2/3. If you ask the guard, the guard will tell you which one of the other prisoners will be released. Your chance of being released apparently drops to 1/2.

Should prisoner A ask the guard? What has gone wrong with his reasoning? This problem is tricky in the sense that the verbal argument of prisoner A seems flawless. If he asked the guard, indeed, the game would be reduced to two people. However, this does not seem correct, because regardless of what the guard says, the probability for A to be pardoned should remain unchanged. Let's see how we can solve this puzzle.

Let X_A, X_B, X_C be the events of sentencing prisoners A, B, C, respectively. Let G_B be the event that the guard says that the prisoner B is released. Without doing anything, we know that

$$\mathbb{P}[X_A] = \frac{1}{3}, \qquad \mathbb{P}[X_B] = \frac{1}{3}, \qquad \mathbb{P}[X_C] = \frac{1}{3}.$$

Conditioned on these events, we can compute the following conditional probabilities that the guard says B is pardoned:

$$\mathbb{P}[G_B \mid X_A] = \frac{1}{2}, \qquad \mathbb{P}[G_B \mid X_B] = 0, \qquad \mathbb{P}[G_B \mid X_C] = 1.$$

Why are these conditional probabilities? $\mathbb{P}[G_B \mid X_B] = 0$ quite straightforward. If the king decides to sentence B, the guard has no way of saying that B will be pardoned. Therefore, $\mathbb{P}[G_B \mid X_B]$ must be zero. $\mathbb{P}[G_B \mid X_C] = 1$ is also not difficult. If the king decides to sentence C, then the guard has no way to tell you that B will be pardoned because the guard cannot say anything about prisoner A. Finally, $\mathbb{P}[G_B \mid X_A] = \frac{1}{2}$ can be understood as follows: If the king decides to sentence A, the guard can either tell you B or C. In other words, the guard flips a coin.

With these conditional probabilities ready, we can determine the probability. This is the conditional probability $\mathbb{P}[X_A \mid G_B]$. That is, supposing that the guard says B is pardoned, what is the probability that A will be sentenced? This is the actual scenario that A is facing. Solving for this conditional probability is not difficult. By Bayes' theorem we know that

$$\mathbb{P}[X_A \mid G_B] = \frac{\mathbb{P}[G_B \mid X_A]\mathbb{P}[X_A]}{\mathbb{P}[G_B]},$$

and $\mathbb{P}[G_B] = \mathbb{P}[G_B|X_A]\mathbb{P}[X_A] + \mathbb{P}[G_B|X_B]\mathbb{P}[X_B] + \mathbb{P}[G_B|X_C]\mathbb{P}[X_C]$ according to the law of total probability. Substituting the numbers into these equations, we have that

$$\begin{aligned}\mathbb{P}[G_B] &= \mathbb{P}[G_B|X_A]\mathbb{P}[X_A] + \mathbb{P}[G_B|X_B]\mathbb{P}[X_B] + \mathbb{P}[G_B|X_C]\mathbb{P}[X_C] \\ &= \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{1}{2}, \\ \mathbb{P}[X_A \mid G_B] &= \frac{\mathbb{P}[G_B \mid X_A]\mathbb{P}[X_A]}{\mathbb{P}[G_B]} = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}.\end{aligned}$$

Therefore, given that the guard says B is pardoned, the probability that A will be sentenced remains $\frac{1}{3}$. In fact, what you can show in this example is that $\mathbb{P}[X_A \mid G_B] = \frac{1}{3} = \mathbb{P}[X_A]$. Therefore, the presence or absence of the guard does not alter the probability. This is because what the guard says is independent of whether the prisoners will be pardoned. The lesson we learn from this problem is not to rely on verbal arguments. We need to write down the conditional probabilities and spell out the steps.

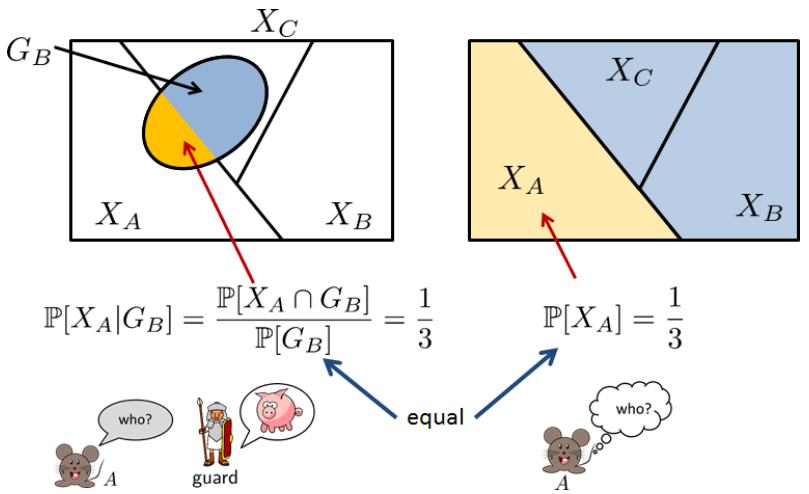


Figure 2.36: The Three Prisoners problem is resolved by noting that $\mathbb{P}[X_A|G_B] = \mathbb{P}[X_A]$. Therefore, the events X_A and G_B are independent.

How to resolve the Three Prisoners problem?

- The key is that G_A, G_B, G_C do not form a **partition**. See **Figure 2.36**.
- $G_B \neq X_B$. When G_B happens, the remaining set is not $X_A \cup X_C$.
- The ratio $\mathbb{P}[X_A \cap G_B]/\mathbb{P}[G_B]$ equals $\mathbb{P}[X_A]$. This is **independence**.

2.5 Summary

By now, we hope that you have become familiar with our slogan **probability is a measure of the size of a set**. Let us summarize:

- **Probability** = a probability law $\mathbb{P}$. You can also view it as the **value** returned by $\mathbb{P}$.
- **Measure** = a ruler, a scale, a stopwatch, or another measuring device. It is a tool that tells you how large or small a set is. The measure has to be compatible with the set. If a set is finite, then the measure can be a counter. If a set is a continuous interval, then the measure can be the length of the interval.
- **Size** = the relative weight of the set for the sample space. Measuring the size is done by using a weighting function. Think of a fair coin versus a biased coin. The former has a uniform weight, whereas the latter has a nonuniform weight.
- **Set** = an event. An event is a subset in the sample space. A probability law $\mathbb{P}$ always maps a **set** to a number. This is different from a typical function that maps a number to another number.

If you understand what this slogan means, you will understand why probability can be applied to discrete events, continuous events, events in n-D spaces, etc. You will also understand the notion of **measure zero** and the notion of **almost sure**. These concepts lie at the foundation of modern data science, in particular, theoretical machine learning.

The second half of this chapter discusses the concept of **conditional probability**. Conditional probability is a metaconcept that can be applied to any measure you use. The motivation of conditional probability is to restrict the probability to a subevent happening in the sample space. If B has happened, the probability for A to *also* happen is $\mathbb{P}[A \cap B]/\mathbb{P}[B]$. If two events are not influencing each other, then we say that A and B are independent. According to Bayes' theorem, we can also switch the order of A given B and B given A, according to Bayes' theorem. Finally, the law of total probability gives us a way to decompose events into subevents.

We end this chapter by mentioning a few terms related to conditional probabilities that will become useful later. Let us use the tennis tournament as an example:

- $\mathbb{P}[W \mid A]$ = **conditional probability** = Given that you played with player A, what is the probability that you will win?
- $\mathbb{P}[A]$ = **prior probability** = Without even entering the game, what is the chance that you will face player A?
- $\mathbb{P}[A \mid W]$ = **posterior probability** = After you have won the game, what is the probability that you have actually played with A?

In many practical engineering problems, the question of interest is often the last one. That is, supposing that you have observed something, what is the most likely cause of that event? For example, supposing we have observed this particular dataset, what is the best Gaussian model that would fit the dataset? Questions like these require some analysis of conditional probability, prior probability, and posterior probability.

2.6 References

Introduction to Probability

2-1 Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd Edition, 2008. Chapter 1.

2-2 Mark D. Ward and Ellen Gundlach, *Introduction to Probability*, W.H. Freeman and Company, 2016. Chapter 1 – Chapter 6.

2-3 Roy D. Yates and David J. Goodman, *Probability and Stochastic Processes*, 3rd Edition, Wiley 2013, Chapter 1.

2-4 John A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, 2006. Chapter 2.

2-5 Sheldon Ross, *A First Course in Probability*, Prentice Hall, 8th Edition, 2010. Chapter 2 and Chapter 3.

2-6 Ani Adhikari and Jim Pitman, *Probability for Data Science*, `http://prob140.org/textbook/content/README.html`. Chapters 1 and 2.

2-7 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Chapter 2.1 – 2.7.

2-8 Athanasios Papoulis and S. Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th Edition, 2001. Chapter 2.

2-9 Henry Stark and John Woods, *Probability and Random Processes With Applications to Signal Processing*, Prentice Hall, 3rd Edition, 2001. Chapter 1.

Measure-Theoretic Probability

2-10 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Chapter 2.8 and 2.9.

2-11 Henry Stark and John Woods, *Probability and Random Processes With Applications to Signal Processing*, Prentice Hall, 3rd Edition, 2001. Appendix D.

2-12 William Feller, *An Introduction to Probability Theory and Its Applications*, Wiley and Sons, 3rd Edition, 1950.

2-13 Andrey Kolmogorov, *Foundations of the Theory of Probability*, 2nd English Edition, Dover 2018. (Translated from Russian to English. Originally published in 1950 by Chelsea Publishing Company New York.)

2-14 Patrick Billingsley, *Probability and Measure*, Wiley, 3rd Edition, 1995.

Real Analysis

2-15 Tom M. Apostol, *Mathematical Analysis*, Pearson, 1974.

2-16 Walter Rudin, *Principles of Mathematical Analysis*, McGraw Hill, 1976.

2.7 Problems

Exercise 1.
A space S and three of its subsets are given by $S = \{1, 3, 5, 7, 9, 11\}$, $A = \{1, 3, 5\}$, $B = \{7, 9, 11\}$, and $C = \{1, 3, 9, 11\}$. Find $A \cap B \cap C$, $A^c \cap B$, $A - C$, and $(A - B) \cup B$.

Exercise 2.
Let $A = (-\infty, r]$ and $B = (-\infty, s]$ where $r \leq s$. Find an expression for $C = (r, s]$ in terms of A and B. Show that $B = A \cup C$, and $A \cap C = \emptyset$.

Exercise 3. (Video Solution)
Simplify the following sets.

(a) $[1, 4] \cap ([0, 2] \cup [3, 5])$

(b) $([0, 1] \cup [2, 3])^c$

(c) $\bigcap_{i=1}^{\infty}(-1/n, +1/n)$

(d) $\bigcup_{i=1}^{\infty}[5, 8 - (2n)^{-1}]$

Exercise 4.
We will sometimes deal with the relationship between two sets. We say that A implies B when A is a subset of B (why?). Show the following results.

(a) Show that if A implies B, and B implies C, then A implies C.

(b) Show that if A implies B, then B^c implies A^c.

Exercise 5.
Show that if $A \cup B = A$ and $A \cap B = A$, then $A = B$.

Exercise 6.
A space S is defined as $S = \{1, 3, 5, 7, 9, 22\}$, and three subsets as $A = \{1, 3, 5\}$, $B = \{7, 9, 11\}$, $C = \{1, 3, 9, 11\}$. Assume that each element has probability $1/6$. Find the following probabilities:

(a) $\mathbb{P}[A]$

(b) $\mathbb{P}[B]$

(c) $\mathbb{P}[C]$

(d) $\mathbb{P}[A \cup B]$

(e) $\mathbb{P}[A \cup C]$

(f) $\mathbb{P}[(A \backslash C) \cup B]$

Exercise 7. (Video Solution)
A collection of 26 letters, a-z, is mixed in a jar. Two letters are drawn at random, one after the other. What is the probability of drawing a vowel (a,e,i,o,u) and a consonant in either order? What is the sample space?

Exercise 8.
Consider an experiment consisting of rolling a die twice. The outcome of this experiment is an ordered pair whose first element is the first value rolled and whose second element is the second value rolled.

(a) Find the sample space.

(b) Find the set A representing the event that the value on the first roll is greater than or equal to the value on the second roll.

(c) Find the set B corresponding to the event that the first roll is a six.

(d) Let C correspond to the event that the first valued rolled and the second value rolled differ by two. Find $A \cap C$.

Note that A, B, and C should be subsets of the sample space specified in Part (a).

Exercise 9.
A pair of dice are rolled.

(a) Find the sample space Ω

(b) Find the probabilities of the events: (i) the sum is even, (ii) the first roll is equal to the second, (iii) the first roll is larger than the second.

Exercise 10.
Let A, B and C be events in an event space. Find expressions for the following:

(a) Exactly one of the three events occurs.

(b) Exactly two of the events occurs.

(c) Two or more of the events occur.

(d) None of the events occur.

Exercise 11.
A system is composed of five components, each of which is either working or failed. Consider an experiment that consists of observing the status of each component, and let the outcomes of the experiment be given by all vectors $(x_1, x_2, x_3, x_4, x_5)$, where x_i is 1 if component i is working and 0 if component i is not working.

(a) How many outcomes are in the sample space of this experiment?

(b) Suppose that the system will work if components 1 and 2 are both working, or if components 3 and 4 are both working, or if components 1, 3, and 5 are all working. Let W be the event that the system will work. Specify all of the outcomes in W.

(c) Let A be the event that components 4 and 5 have both failed. How many outcomes are in the event A?

(d) Write out all outcomes in the event $A \cap W$.

Exercise 12. (Video Solution)
A number x is selected at random in the interval $[-1, 2]$. Let the events $A = \{x \,|\, x < 0\}$, $B = \{x \,|\, |x - 0.5| < 0.5\}$, $C = \{x \,|\, x > 0.75\}$. Find (a) $\mathbb{P}[A \,|\, B]$, (b) $\mathbb{P}[B \,|\, C]$, (c) $\mathbb{P}[A \,|\, C^c]$, (d) $\mathbb{P}[B \,|\, C^c]$.

Exercise 13. (Video Solution)
Let the events A and B have $\mathbb{P}[A] = x$, $\mathbb{P}[B] = y$ and $\mathbb{P}[A \cup B] = z$. Find the following probabilities: (a) $\mathbb{P}[A \cap B]$, (b) $\mathbb{P}[A^c \cap B^c]$, (c) $\mathbb{P}[A^c \cup B^c]$, (d) $\mathbb{P}[A \cap B^c]$, (e) $\mathbb{P}[A^c \cup B]$.

Exercise 14.

(a) By using the fact that $\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$, show that $\mathbb{P}[A \cup B \cup C] \leq \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[C]$.

(b) By using the fact that $\mathbb{P}\left[\bigcup_{k=1}^{n} A_k\right] \leq \sum_{k=1}^{n} \mathbb{P}[A_k]$, show that

$$\mathbb{P}\left[\bigcap_{k=1}^{n} A_k\right] \geq 1 - \sum_{k=1}^{n} \mathbb{P}[A_k^c].$$

Exercise 15.
Use the distributive property of set operations to prove the following generalized distributive law:

$$A \cup \left(\bigcap_{i=1}^{n} B_i\right) = \bigcap_{i=1}^{n} (A \cup B_i).$$

Hint: Use mathematical induction. That is, show that the above is true for $n = 2$ and that it is also true for $n = k + 1$ when it is true for $n = k$.

Exercise 16.
The following result is known as the Bonferroni's Inequality.

(a) Prove that for any two events A and B, we have

$$\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1.$$

(b) Generalize the above to the case of n events $A_1, A_2, \ldots, A_n$, by showing that

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) \geq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n) - (n - 1).$$

Hint: You may use the generalized Union Bound $\mathbb{P}(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$.

Exercise 17. (Video Solution)
Let A, B, C be events with probabilities $\mathbb{P}[A] = 0.5$, $\mathbb{P}[B] = 0.2$, $\mathbb{P}[C] = 0.4$. Find

(a) $\mathbb{P}[A \cup B]$ if A and B are independent.

(b) $\mathbb{P}[A \cup B]$ if A and B are disjoint.

(c) $\mathbb{P}[A \cup B \cup C]$ if A, B and C are independent.

(d) $\mathbb{P}[A \cup B \cup C]$ if A, B and C are pairwise disjoint; can this happen?

Exercise 18. (Video Solution)
A block of information is transmitted repeated over a noisy channel until an error-free block is received. Let $M \geq 1$ be the number of blocks required for a transmission. Define the following sets.

(i) $A = \{M \text{ is even}\}$

(ii) $B = \{M \text{ is a multiple of } 3\}$

(iii) $C = \{M \text{ is less than or equal to } 6\}$

Assume that the probability of requiring one additional block is half of the probability without the additional block. That is:

$$\mathbb{P}[M = k] = \left(\frac{1}{2}\right)^k, \qquad k = 1, 2, \ldots.$$

Determine the following probabilities.

(a) $\mathbb{P}[A]$, $\mathbb{P}[B]$, $\mathbb{P}[C]$, $\mathbb{P}[C^c]$

(b) $\mathbb{P}[A \cap B]$, $\mathbb{P}[A \backslash B]$, $\mathbb{P}[A \cap B \cap C]$

(c) $\mathbb{P}[A \,|\, B]$, $\mathbb{P}[B \,|\, A]$

(d) $\mathbb{P}[A \,|\, B \cap C]$, $\mathbb{P}[A \cap B \,|\, C]$

Exercise 19. (Video Solution)
A binary communication system transmits a signal X that is either a $+2$-voltage signal or a -2-voltage signal. A malicious channel reduces the magnitude of the received signal by the number of heads it counts in two tosses of a coin. Let Y be the resulting signal. Possible values of Y are listed below.

	2 Heads	1 Head	No Head
$X = -2$	$Y = 0$	$Y = -1$	$Y = -2$
$X = +2$	$Y = 0$	$Y = +1$	$Y = +2$

Assume that the probability of having $X = +2$ and $X = -2$ is equal.

(a) Find the sample space of Y, and hence the probability of each value of Y.

(b) What are the probabilities $\mathbb{P}[X = +2 \,|\, Y = 1]$ and $\mathbb{P}[Y = 1 \,|\, X = -2]$?

Exercise 20. (Video Solution)
A block of 100 bits is transmitted over a binary communication channel with a probability of bit error $p = 10^{-2}$.

(a) If the block has 1 or fewer errors, then the receiver accepts the block. Find the probability that the block is accepted.

(b) If the block has more than 1 error, then the block is retransmitted. What is the probability that 4 blocks are transmitted?

Exercise 21. (Video Solution)
A machine makes errors in a certain operation with probability p. There are two types of errors. The fraction of errors that are type A is α and the fraction that are type B is $1-\alpha$.

(a) What is the probability of k errors in n operations?

(b) What is the probability of k_1 type A errors in n operations?

(c) What is the probability of k_2 type B errors in n operations?

(d) What is the joint probability of k_1 type A errors and k_2 type B errors in n operations? Hint: There are $\binom{n}{k_1}\binom{n-k_1}{k_2}$ possibilities of having k_1 type A errors and k_2 type B errors in n operations. (Why?)

Exercise 22. (Video Solution)
A computer manufacturer uses chips from three sources. Chips from sources A, B and C are defective with probabilities 0.005, 0.001 and 0.01, respectively. The proportions of chips from A, B and C are 0.5, 0.1 and 0.4 respectively. If a randomly selected chip is found to be defective, find

(a) the probability that the chips are from A.

(b) the probability that the chips are from B.

(c) the probability that the chips are from C.

Exercise 23. (Video Solution)
In a lot of 100 items, 50 items are defective. Suppose that m items are selected for testing. We say that the manufacturing process is malfunctioning if the probability that one or more items are tested to be defective. Call this failure probability p. What should be the minimum m such that $p \geq 0.99$?

Exercise 24. (Video Solution)
One of two coins is selected at random and tossed three times. The first coin comes up heads with probability $p_1 = 1/3$ and the second coin with probability $p_2 = 2/3$.

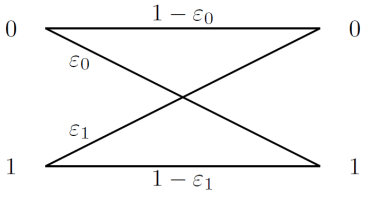
(a) What is the probability that the number of heads is $k = 3$?

(b) Repeat (a) for $k = 0, 1, 2$.

(c) Find the probability that coin 1 was tossed given that k heads were observed, for $k = 0, 1, 2, 3$.

(d) In part (c), which coin is more probably when 2 heads have been observed?

Exercise 25. (Video Solution)
Consider the following communication channel. A source transmits a string of binary symbols through a noisy communication channel. Each symbol is 0 or 1 with probability p and $1-p$, respectively, and is received incorrectly with probability ε_0 and ε_1. Errors in different symbols transmissions are independent.



Denote S as the source and R as the receiver.

(a) What is the probability that a symbol is correctly received? Hint: Find

$$\mathbb{P}[R=1 \cap S=1] \quad \text{and} \quad \mathbb{P}[R=0 \cap S=0].$$

(b) Find the probability of receiving 1011 conditioned on that 1011 was sent, i.e.,

$$\mathbb{P}[R=1011 \,|\, S=1011].$$

(c) To improve reliability, each symbol is transmitted three times, and the received string is decoded by the majority rule. In other words, a 0 (or 1) is transmitted as 000 (or 111, respectively), and it is decoded at the receiver as a 0 (or 1) if and only if the received three-symbol string contains at least two 0s (or 1s, respectively). What is the probability that the symbol is correctly decoded, given that we send a 0?

(d) Suppose that the scheme of part (c) is used. What is the probability that a 0 was sent if the string 101 was received?

(e) Suppose the scheme of part (c) is used and given that a 0 was sent. For what value of ε_0 is there an improvement in the probability of correct decoding? Assume that $\varepsilon_0 \neq 0$.

Chapter 3

Discrete Random Variables

When working on a data analysis problem, one of the biggest challenges is the disparity between the theoretical tools we learn in school and the *actual data* our boss hands to us. By actual data, we mean a collection of numbers, perhaps organized or perhaps not. When we are given the dataset, the first thing we do would certainly not be to define the Borel σ-field and then define the measure. Instead, we would normally compute the mean, the standard deviation, and perhaps some scores about the skewness.

The situation is best explained by the landscape shown in **Figure 3.1**. On the one hand, we have well-defined probability tools, but on the other hand, we have a set of practical "battle skills" for processing data. Often we view them as two separate entities. As long as we can pull the statistics from the dataset, why bother about the theory? Alternatively, we have a set of theories, but we will never verify them using the actual datasets. How can we bridge the two? What are the missing steps in the probability theory we have learned so far? The goal of this chapter (and the next) is to fill this gap.

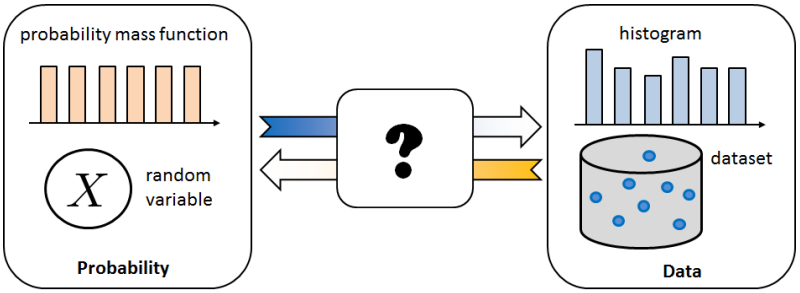


Figure 3.1: The landscape of probability and data. Often we view probability and data analysis as two different entities. However, probability and data analysis are inseparable. The goal of this chapter is to link the two.

Three concepts to bridge the gap between theory and practice

The starting point of our discussion is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. It is an abstract concept, but we hope we have convinced you in Chapter 2 of its significance. However, the probability space is certainly not "user friendly" because no one would write a Python program to

implement those theories. How do we make the abstract probability space more convenient so that we can model practical scenarios?

The first step is to recognize that the sample space and the event space are all based on *statements*, for example, "getting a head when flipping a coin" or "winning the game." These statements are not numbers, but we (engineers) love numbers. Therefore, we should ask a very basic question: How do we convert a statement to a number? The answer is the concept of **random variables**.

Key Concept 1: What are random variables?

Random variables are mappings from events to numbers.

Now, suppose that we have constructed a random variable that translates statements to numbers. The next task is to endow the random variable with probabilities. More precisely, we need to assign probabilities to the random variable so that we can perform computations. This is done using the concept called **probability mass function** (PMF).

Key Concept 2: What are probability mass functions (PMFs)?

Probability mass functions are the ideal histograms of random variables.

The best way to think about a PMF is a histogram, something we are familiar with. A histogram has two axes: The x-axis denotes the set of **states** and the y-axis denotes the **probability**. For each of the states that the random variable possesses, the histogram tells us the probability of getting a particular state. The PMF is the *ideal* histogram of a random variable. It provides a complete characterization of the random variable. If you have a random variable, you must specify its PMF. Vice versa, if you tell us the PMF, you have specified a random variable.

We ask the third question about pulling information from the probability mass function, such as the mean and standard deviation. How do we obtain these numbers from the PMF? We are also interested in operations on the mean and standard deviations. For example, if a professor offers ten bonus points to the entire class, how will it affect the mean and standard deviation? If a store provides 20% off on all its products, what will happen to its mean retail price and standard deviation? However, the biggest question is perhaps the difference between the mean we obtain from a PMF and the mean we obtain from a histogram. Understanding this difference will immediately help us build a bridge from theory to practice.

Key Concept 3: What is expectation?

Expectation = Mean = Average computed from a PMF.

Organization of this chapter

The plan for this chapter is as follows. We will start with the basic concepts of random variables in Section 3.1. We will formally define the random variables and discuss their relationship with the abstract probability space. Once this linkage is built, we can put

the abstract probability space aside and focus on the random variables. In Section 3.2 we will define the probability mass function (PMF) of a random variable, which tells us the probability of obtaining a state of the random variable. PMF is closely related to the histogram of a dataset. We will explain the connection. In Section 3.3 we take a small detour to consider the cumulative distribution functions (CDF). Then, we discuss the mean and standard deviation in Section 3.4. Section 3.5 details a few commonly used random variables, including Bernoulli, binomial, geometric, and Poisson variables.

3.1 Random Variables

3.1.1 A motivating example

Consider an experiment with 4 outcomes $\Omega = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$. We want to construct the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sample space Ω is already defined. The event space $\mathcal{F}$ is the set of all possible subsets in Ω, which, in our case, is a set of 2^4 subsets. For the probability law $\mathbb{P}$, let us assume that the probability of obtaining each outcome is

$$\mathbb{P}[\{\clubsuit\}] = \frac{1}{6}, \quad \mathbb{P}[\{\diamondsuit\}] = \frac{2}{6}, \quad \mathbb{P}[\{\heartsuit\}] = \frac{2}{6}, \quad \mathbb{P}[\{\spadesuit\}] = \frac{1}{6}.$$

Therefore, we have constructed a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where everything is perfectly defined. So, in principle, they can live together happily forever.

A lazy data scientist comes, and there is a (small) problem. The data scientist does not want to write the symbols $\clubsuit, \diamondsuit, \heartsuit, \spadesuit$. There is nothing wrong with his motivation because all of us want efficiency. How can we help him? Well, the easiest solution is to *encode* each symbol with a number, for example, $\clubsuit \leftarrow 1$, $\diamondsuit \leftarrow 2$, $\heartsuit \leftarrow 3$, $\spadesuit \leftarrow 4$, where the arrow means that we assign a number to the symbol. But we can express this more formally by defining a function $X : \Omega \to \mathbb{R}$ with

$$X(\clubsuit) = 1, \quad X(\diamondsuit) = 2, \quad X(\heartsuit) = 3, \quad X(\spadesuit) = 4.$$

There is nothing new here: we have merely converted the symbols to numbers, with the help of a function X. However, with X defined, the probabilities can be written as

$$\mathbb{P}[X = 1] = \frac{1}{6}, \quad \mathbb{P}[X = 2] = \frac{2}{6}, \quad \mathbb{P}[X = 3] = \frac{2}{6}, \quad \mathbb{P}[X = 4] = \frac{1}{6}.$$

This is much more convenient, and so the data scientist is happy.

3.1.2 Definition of a random variable

The story above is exactly the motivation for random variables. Let us define a random variable formally.

Definition 3.1. *A* **random variable** X *is a function* $X : \Omega \to \mathbb{R}$ *that maps an outcome* $\xi \in \Omega$ *to a number* $X(\xi)$ *on the real line.*

This definition may be puzzling at first glance. Why should we overcomplicate things by defining a *function* and calling it a *variable*?

If you recall the story above, we can map the notations of the story to the notations of the definition as follows.

Symbol	Meaning
Ω	sample space = the set containing $\clubsuit, \diamondsuit, \heartsuit, \spadesuit$
ξ	an element in the sample space, which is one of $\clubsuit, \diamondsuit, \heartsuit, \spadesuit$
X	a function that maps $\clubsuit$ to the number 1, $\diamondsuit$ to the number 2, etc
$X(\xi)$	a number on the real line, e.g., $X(\clubsuit) = 1$

This explains our informal definition of random variables:

Key Concept 1: What are random variables?

Random variables are mappings from events to numbers.

The random variable X is a *function*. The input to the function is an outcome of the sample space, whereas the output is a number on the real line. This type of function is somewhat different from an ordinary function that often translates a number to another number. Nevertheless, X is a function.

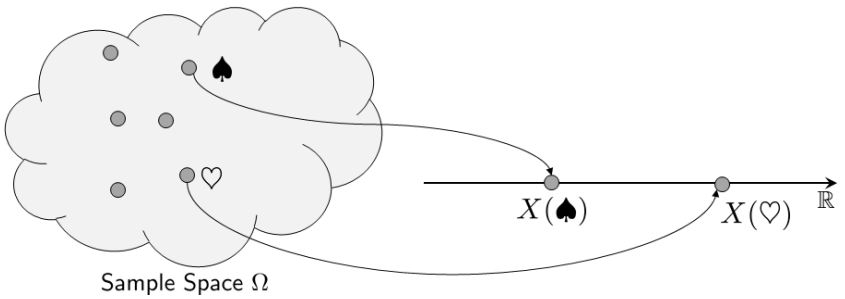


Figure 3.2: A random variable is a mapping from the outcomes in the sample space to numbers on the real line. We can think of a random variable X as a translator that translates a statement to a number.

Why do we call this function X a *variable*? **X is a variable because X has multiple states**. As we illustrate in **Figure 3.2**, the mapping X translates every outcome ξ to a number. There are multiple numbers, which are the states of X. Each state has a certain probability for X to land on. Because X is not deterministic, we call it a *random* variable.

Example 3.1. Suppose we flip a fair coin so that $\Omega = \{\text{head}, \text{tail}\}$. We can define the random variable $X : \Omega \to \mathbb{R}$ as

$$X(\text{head}) = 1, \qquad \text{and} \qquad X(\text{tail}) = 0.$$

Therefore, when we write $\mathbb{P}[X = 1]$ we actually mean $\mathbb{P}[\{\text{head}\}]$. Is there any difference between $\mathbb{P}[\{\text{Head}\}]$ and $\mathbb{P}[X = 1]$? No, because they are describing two identical events. Note that the assignment of the value is totally up to you. You can say "head" is equal to the value 102. This is allowed and legitimate, but it isn't very convenient.

Example 3.2. Flip a coin 2 times. The sample space Ω is

$$\Omega = \{(\text{head}, \text{head}), (\text{head}, \text{tail}), (\text{tail}, \text{head}), (\text{tail}, \text{tail})\}.$$

Suppose that X is a random variable that maps an outcome to a number representing the sum of "head," i.e.,

$$X(\cdot) = \text{number of heads}.$$

Then, for the 4 ξ's in the sample space there are only 3 distinct numbers. More precisely, if we let $\xi_1 = (\text{head}, \text{head})$, $\xi_2 = (\text{head}, \text{tail})$, $\xi_3 = (\text{tail}, \text{head})$, $\xi_4 = (\text{tail}, \text{tail})$, then, we have

$$X(\xi_1) = 2, \quad X(\xi_2) = 1, \quad X(\xi_3) = 1, \quad X(\xi_4) = 0.$$

A pictorial illustration of this random variable is shown in **Figure 3.3**. This example shows that the mapping defined by the random variable is not necessarily a one-to-one mapping because multiple outcomes can be mapped to the same number.

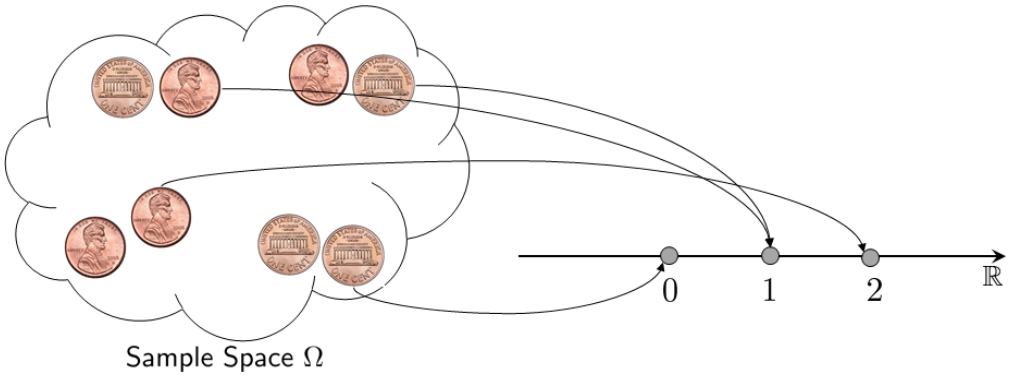


Figure 3.3: A random variable that maps a pair of coins to a number, where the number represents the number of heads.

3.1.3 Probability measure on random variables

By now, we hope that you understand Key Concept 1: **A random variable is a mapping from a statement to a number**. However, we are now facing another difficulty. We knew how to measure the size of an event using the probability law $\mathbb{P}$ because $\mathbb{P}(\cdot)$ takes an event $E \in \mathcal{F}$ and sends it to a number between $[0, 1]$. After the translation X, we cannot send the output $X(\xi)$ to $\mathbb{P}(\cdot)$ because $\mathbb{P}(\cdot)$ "eats" a set $E \in \mathcal{F}$ and not a number $X(\xi) \in \mathbb{R}$. Therefore, when we write $\mathbb{P}[X = 1]$, how do we measure the size of the event $X = 1$?

This question appears difficult but is actually quite easy to answer. Since the probability law $\mathbb{P}(\cdot)$ is always applied to an **event**, we need to define an event for the random variable X. If we write the sets clearly, we note that "$X = a$" is equivalent to the set

$$E = \left\{ \xi \in \Omega \,\middle|\, X(\xi) = a \right\}.$$

This is the set that contains all possible ξ's such that $X(\xi) = a$. Therefore, when we say "find the probability of $X = a$," we are effectively asking the size of the set $E = \{\xi \in \Omega \,|\, X(\xi) = a\}$.

How then do we measure the size of E? Since E is a subset in the sample space, E is measurable by $\mathbb{P}$. All we need to do is to determine what E is for a given a. This, in turn, requires us to find the **pre-image** $X^{-1}(a)$, which is defined as

$$X^{-1}(a) \stackrel{\text{def}}{=} \left\{ \xi \in \Omega \,\middle|\, X(\xi) = a \right\}.$$

Wait a minute, is this set just equal to E? Yes, the event E we are seeking is exactly the pre-image $X^{-1}(a)$. As such, the probability measure of E is

$$\mathbb{P}[X = a] = \mathbb{P}[X^{-1}(a)].$$

Figure 3.4 illustrates a situation where two outcomes ξ_1 and ξ_2 are mapped to the same value a on the real line. The corresponding event is the set $X^{-1}(a) = \{\xi_1, \xi_2\}$.

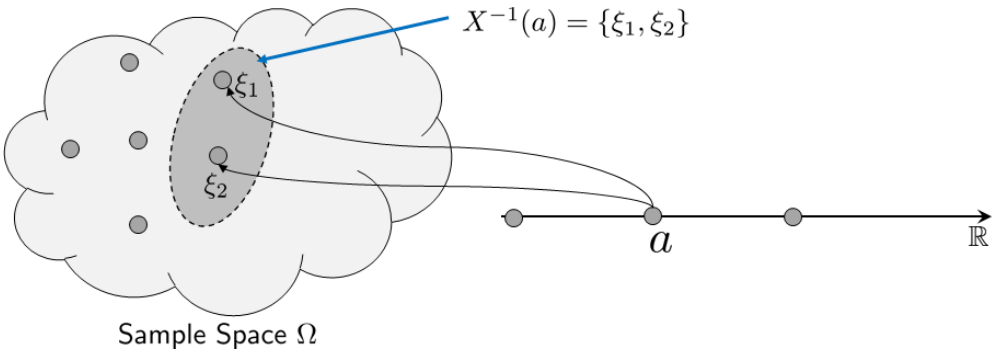


Figure 3.4: When computing the probability of $\mathbb{P}[\{\xi \in \Omega \,|\, X(\xi) = a\}]$, we effectively take the inverse mapping $X^{-1}(a)$ and compute the probability of the event $\mathbb{P}[\{\xi \in X^{-1}(a)\}] = \mathbb{P}[\{\xi_1, \xi_2\}]$.

Example 3.3. Suppose we throw a die. The sample space is

$$\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}.$$

There is a natural mapping X that maps $X(\boxdot) = 1$, $X(\boxdot) = 2$ and so on. Thus,

$$\begin{aligned}\mathbb{P}[X \le 3] &\overset{(a)}{=} \mathbb{P}[X=1] + \mathbb{P}[X=2] + \mathbb{P}[X=3] \\ &\overset{(b)}{=} \mathbb{P}[X^{-1}(1)] + \mathbb{P}[X^{-1}(2)] + \mathbb{P}[X^{-1}(3)] \\ &\overset{(c)}{=} \mathbb{P}[\{\boxdot\}] + \mathbb{P}[\{\boxdot\}] + \mathbb{P}[\{\boxdot\}] = \frac{3}{6}.\end{aligned}$$

In this derivation, step (a) is based on Axiom III, where the three events are disjoint. Step (b) is the pre-image due to the random variable X. Step (c) is the list of actual events in the event space. Note that there is no hand-waving argument in this derivation. Every step is justified by the concepts and theorems we have learned so far.

Example 3.4. Throw a die twice. The sample space is then

$$\Omega = \{(\boxdot, \boxdot), (\boxdot, \boxdot), \ldots, (\boxdot, \boxdot)\}.$$

These elements can be translated to 36 outcomes:

$$\xi_1 = (\boxdot, \boxdot), \xi_2 = (\boxdot, \boxdot), \ldots, \xi_{36} = (\boxdot, \boxdot).$$

Let

$$X = \text{sum of two numbers.}$$

Then, if we want to find the probability of getting $X = 7$, we can trace back and ask: Among the 36 outcomes, which of those ξ_i's will give us $X(\xi) = 7$? Or, what is the set $X^{-1}(7)$? To this end, we can write

$$\begin{aligned}\mathbb{P}[X=7] &= \mathbb{P}[\{(\boxdot, \boxdot), (\boxdot, \boxdot), (\boxdot, \boxdot), (\boxdot, \boxdot), (\boxdot, \boxdot), (\boxdot, \boxdot)\}] \\ &= \mathbb{P}[(\boxdot, \boxdot)] + \mathbb{P}[(\boxdot, \boxdot)] + \mathbb{P}[(\boxdot, \boxdot)] \\ &\quad + \mathbb{P}[(\boxdot, \boxdot)] + \mathbb{P}[(\boxdot, \boxdot)] + \mathbb{P}[(\boxdot, \boxdot)] \\ &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{6}.\end{aligned}$$

Again, in this example, you can see that all the steps are fully justified by the concepts we have learned so far.

Closing remark. In practice, when the problem is clearly defined, we can skip the inverse mapping $X^{-1}(a)$. However, this does not mean that the probability triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is gone; it is still present. The triplet is now just the background of the problem.

The set of all possible values returned by X is denoted as $X(\Omega)$. Since X is not necessarily a bijection, the size of $X(\Omega)$ is not necessarily the same as the size of Ω. The elements in $X(\Omega)$ are often denoted as a or x. We call a or x one of the **states** of X. Be careful not to confuse x and X. The variable X is the random variable; it is a function. The variable x is a state assigned by X. A random variable X has multiple states. When we write $\mathbb{P}[X = x]$, we describe the probability of a random variable X taking a particular state x. It is exactly the same as $\mathbb{P}[\{\xi \in \Omega \,|\, X(\xi) = x\}]$.

3.2 Probability Mass Function

Random variables are mappings that translate events to numbers. After the translation, we have a set of numbers denoting the **states** of the random variables. Each state has a different probability of occurring. The probabilities are summarized by a function known as the probability mass function (PMF).

3.2.1 Definition of probability mass function

Definition 3.2. *The* **probability mass function (PMF)** *of a random variable X is a function which specifies the probability of obtaining a number $X(\xi) = x$. We denote a PMF as*

$$p_X(x) = \mathbb{P}[X = x]. \tag{3.1}$$

The set of all possible states of X is denoted as $X(\Omega)$.

Do not get confused by the sample space Ω and the set of states $X(\Omega)$. The sample space Ω contains all the possible outcomes of the experiments, whereas $X(\Omega)$ is the translation by the mapping X. The event $X = a$ is the set $X^{-1}(a) \subseteq \Omega$. Therefore, when we say $\mathbb{P}[X = x]$ we really mean $\mathbb{P}[X^{-1}(x)]$.

The probability mass function is a histogram summarizing the probability of each of the states X takes. Since it is a histogram, a PMF can be easily drawn as a bar chart.

Example 3.5. Flip a coin twice. The sample space is $\Omega = \{\text{HH, HT, TH, TT}\}$. We can assign a random variable X = number of heads. Therefore,

$$X(\text{“HH”}) = 2, X(\text{“TH”}) = 1, X(\text{“HT”}) = 1, X(\text{“TT”}) = 0.$$

So the random variable X takes three states: 0, 1, 2. The PMF is therefore

$$\begin{aligned} p_X(0) &= \mathbb{P}[X = 0] = \mathbb{P}[\{\text{“TT”}\}] = \frac{1}{4}, \\ p_X(1) &= \mathbb{P}[X = 1] = \mathbb{P}[\{\text{“TH”}, \text{“HT”}\}] = \frac{1}{2}, \\ p_X(2) &= \mathbb{P}[X = 2] = \mathbb{P}[\{\text{“HH”}\}] = \frac{1}{4}. \end{aligned}$$

3.2.2 PMF and probability measure

In Chapter 2, we learned that probability is a measure of the size of a set. We introduced a **weighting function** that weights each of the elements in the set. The PMF is the weighing function for discrete random variables. Two random variables are different when their PMFs are different because they are constructing two different measures.

To illustrate the idea, suppose there are two dice. They each have probability masses as follows.

$$\mathbb{P}[\{\boxdot\}] = \frac{1}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{3}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{4}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{1}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{1}{12},$$
$$\mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12},\ \mathbb{P}[\{\boxdot\}] = \frac{2}{12},$$

Let us define two random variables, X and Y, for the two dice. Then, the PMFs p_X and p_Y can be defined as

$$p_X(1) = \frac{1}{12},\ p_X(2) = \frac{2}{12},\ p_X(3) = \frac{3}{12},\ p_X(4) = \frac{4}{12},\ p_X(5) = \frac{1}{12},\ p_X(6) = \frac{1}{12},$$
$$p_Y(1) = \frac{2}{12},\ p_Y(2) = \frac{2}{12},\ p_Y(3) = \frac{2}{12},\ p_Y(4) = \frac{2}{12},\ p_Y(5) = \frac{2}{12},\ p_Y(6) = \frac{2}{12}.$$

These two probability mass functions correspond to two different probability measures, let's say $\mathbb{F}$ and $\mathbb{G}$. Define the event $E = \{\text{between 2 and 3}\}$. Then, $\mathbb{F}(E)$ and $\mathbb{G}(E)$ will lead to two different results:

$$\mathbb{F}(E) = \mathbb{P}[2 \le X \le 3] = p_X(2) + p_X(3) = \frac{1}{12} + \frac{2}{12} = \frac{3}{12},$$
$$\mathbb{G}(E) = \mathbb{P}[2 \le Y \le 3] = p_Y(2) + p_Y(3) = \frac{2}{12} + \frac{2}{12} = \frac{4}{12}.$$

Note that even though for some particular events two final results could be the same (e.g., $2 \le X \le 4$ and $2 \le Y \le 4$), the underlying measures are completely different.

Figure 3.5 shows another example of two different measures $\mathbb{F}$ and $\mathbb{G}$ on the same sample space $\Omega = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$. Since the PMFs of the two measures are different, even when given the same event E, the resulting probabilities will be different.

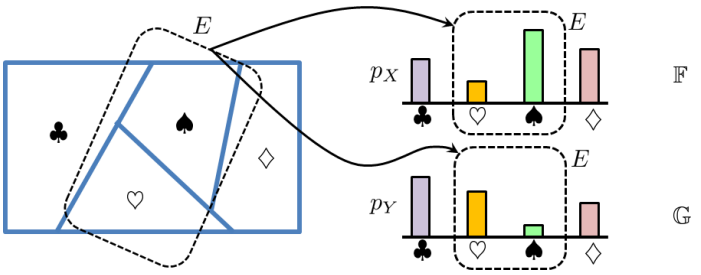


Figure 3.5: If we want to measure the size of a set E, using two different PMFs is equivalent to using two different measures. Therefore, the probabilities will be different.

Does $p_X = p_Y$ imply $X = Y$? If two random variables X and Y have the same PMF, does it mean that the random variables are the same? The answer is no. Consider a random variable with a symmetric PMF, e.g.,

$$p_X(-1) = \frac{1}{4}, \quad p_X(0) = \frac{1}{2}, \quad p_X(1) = \frac{1}{4}. \tag{3.2}$$

Suppose $Y = -X$. Then, $p_Y(-1) = \frac{1}{4}$, $p_Y(0) = \frac{1}{2}$, and $p_Y(1) = \frac{1}{4}$, which is the same as p_X. However, X and Y are two different random variables. If the sample space is $\{\clubsuit, \diamondsuit, \heartsuit\}$, we can define the mappings $X(\cdot)$ and $Y(\cdot)$ as

$$X(\clubsuit) = -1, \quad X(\diamondsuit) = 0, \quad X(\heartsuit) = +1,$$
$$Y(\clubsuit) = +1, \quad Y(\diamondsuit) = 0, \quad Y(\heartsuit) = -1.$$

Therefore, when we say $p_X(-1) = \frac{1}{4}$, the underlying event is ♣. But when we say $p_Y(-1) = \frac{1}{4}$, the underlying event is ♡. The two random variables are different, although their PMFs have exactly the same shape.

3.2.3 Normalization property

Here we must mention one important property of a probability mass function. This property is known as the **normalization property**, which is a useful tool for a sanity check.

Theorem 3.1. *A PMF should satisfy the condition that*

$$\sum_{x \in X(\Omega)} p_X(x) = 1. \tag{3.3}$$

Proof. The proof follows directly from Axiom II, which states that $\mathbb{P}[\Omega] = 1$. Since x covers all numerical values X can take, and since each x is distinct, by Axiom III we have

$$\begin{aligned}\sum_{x \in X(\Omega)} \mathbb{P}[X = x] &= \sum_{x \in X(\Omega)} \mathbb{P}\left[\{\xi \in \Omega \,|\, X(\xi) = x\}\right] \\ &= \mathbb{P}\left[\bigcup_{\xi \in \Omega} \{\xi \in \Omega \,|\, X(\xi) = x\}\right] = \mathbb{P}[\Omega] = 1.\end{aligned}$$

□

Practice Exercise 3.1. Let $p_X(k) = c\left(\frac{1}{2}\right)^k$, where $k = 1, 2, \ldots$. Find c.

Solution. Since $\sum_{k \in X(\Omega)} p_X(k) = 1$, we must have

$$\sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^k = 1.$$

Evaluating the geometric series on the right-hand side, we can show that

$$\begin{aligned}\sum_{k=1}^{\infty} c\left(\frac{1}{2}\right)^k &= \frac{c}{2} \sum_{k=0}^{\infty} \left(\frac{1}{2}\right)^k \\ &= \frac{c}{2} \cdot \frac{1}{1 - \frac{1}{2}} \\ &= c \quad \Longrightarrow \quad c = 1.\end{aligned}$$

Practice Exercise 3.2. Let $p_X(k) = c \cdot \sin\left(\frac{\pi}{2}k\right)$, where $k = 1, 2, \ldots$. Find c.

Solution. The reader may might be tempted to sum $p_X(k)$ over all the possible k's:

$$\sum_{k=1}^{\infty} \sin\left(\frac{\pi}{2}k\right) = 1 + 0 - 1 + 0 + \cdots \stackrel{?}{=} 0.$$

However, a more careful inspection reveals that $p_X(k)$ is actually negative when $k = 3, 7, 11, \ldots$. This cannot happen because a probability mass function must be non-negative. Therefore, the problem is not defined, and so there is no solution.

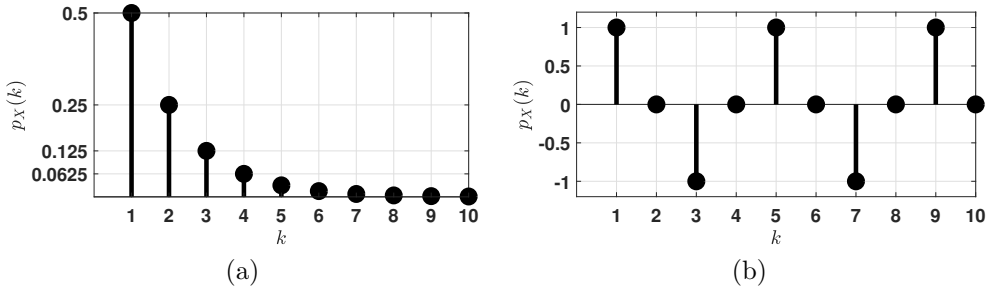


Figure 3.6: (a) The PMF of $p_X(k) = c\left(\frac{1}{2}\right)^k$, for $k = 1, 2, \ldots$. (b) The PMF of $p_X(k) = \sin\left(\frac{\pi}{2}k\right)$, where $k = 1, 2, \ldots$. Note that this is not a valid PMF because probability cannot have negative values.

3.2.4 PMF versus histogram

PMFs are closely related to histograms. A histogram is a plot that shows the frequency of a state. As we see in **Figure 3.6**, the x-axis is a collection of states, whereas the y-axis is the frequency. So a PMF is indeed a histogram.

Viewing a PMF as a histogram can help us understand a random variable. For better or worse, treating a random variable as a histogram could help you differentiate a random variable from a variable. An ordinary variable only has one state, but a random variable has multiple states. At any particular instance, we do not know which state will show up before our observation. However, we do know the probability. For example, in the coin-flip example, while we do not know whether we will get "HH," we know that the chance of getting "HH" is 1/4. Of course, having a probability of 1/4 does not mean that we will get "HH" once every four trials. It only means that if we run an infinite number of experiments, then 1/4 of the experiments will give us "HH."

The linkage between PMF and histogram can be quite practical. For example, while we do not know the true underlying distribution of the 26 letters of the English alphabet, we can collect a large number of words and plot the histogram. The example below illustrates how we can empirically define a random variable from the data.

Example. There are 26 English letters, but the frequencies of the letters in writing are different. If we define a random variable X as a letter we randomly draw from an English text, we can think of X as an object with 26 different states. The mapping associated with the random variable is straightforward: $X(\text{"a"}) = 1$, $X(\text{"b"}) = 2$, etc. The probability of landing on a particular state approximately follows a histogram shown in **Figure 3.7**. The histogram provides meaningful values of the probabilities, e.g., $p_X(1) = 0.0847$, $p_X(2) = 0.0149$, etc. The true probability of the states may not be exactly these values. However, when we have enough samples, we generally expect the histogram to approach the theoretical PMF. The MATLAB and Python codes used to generate this histogram are shown below.

```
% MATLAB code to generate the histogram
load('ch3_data_English');
bar(f/100,'FaceColor',[0.9,0.6,0.0]);
```

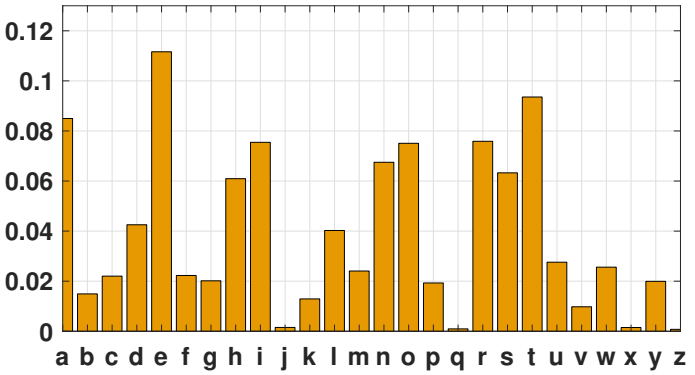


Figure 3.7: The frequency of the 26 English letters. Data source: Wikipedia.

```
xticklabels({'a','b','c','d','e','f','g','h','i','j','k','l',...
    'm','n','o','p','q','r','s','t','u','v','w','x','y','z'});
xticks(1:26);
yticks(0:0.02:0.2);
axis([1 26 0 0.13]);
```

```
# Python code generate the histogram
import numpy as np
import matplotlib.pyplot as plt
f = np.loadtxt('./ch3_data_english.txt')
n = np.arange(26)
plt.bar(n, f/100)
ntag = ['a','b','c','d','e','f','g','h','i','j','k','l','m',...
    'n','o','p','q','r','s','t','u','v','w','x','y','z']
plt.xticks(n, ntag)
```

PMF = ideal histograms

If a random variable is more or less a histogram, why is the PMF such an important concept? The answer to this question has two parts. The first part is that the histogram generated from a dataset is always an **empirical** histogram, so-called because the dataset comes from observation or experience rather than theory. Thus the histograms may vary slightly every time we collect a dataset.

As we increase the number of data points in a dataset, the histogram will eventually converge to an **ideal** histogram, or a **distribution**. For example, counting the number of heads in 100 coin flips will fluctuate more in percentage terms than counting the heads in 10 million coin flips. The latter will almost certainly have a histogram that is closer to a 50–50 distribution. Therefore, the "histogram" generated by a random variable can be considered the ultimate histogram or the limiting histogram of the experiment.

To help you visualize the difference between a PMF and a histogram, we show in **Figure 3.8** an experiment in which a die is thrown N times. Assuming that the die is fair, the PMF is simply $p_X(k) = 1/6$ for $k = 1, \ldots, 6$, which is a uniform distribution across the 6 states. Now, we can throw the die many times. As N increases, we observe that the

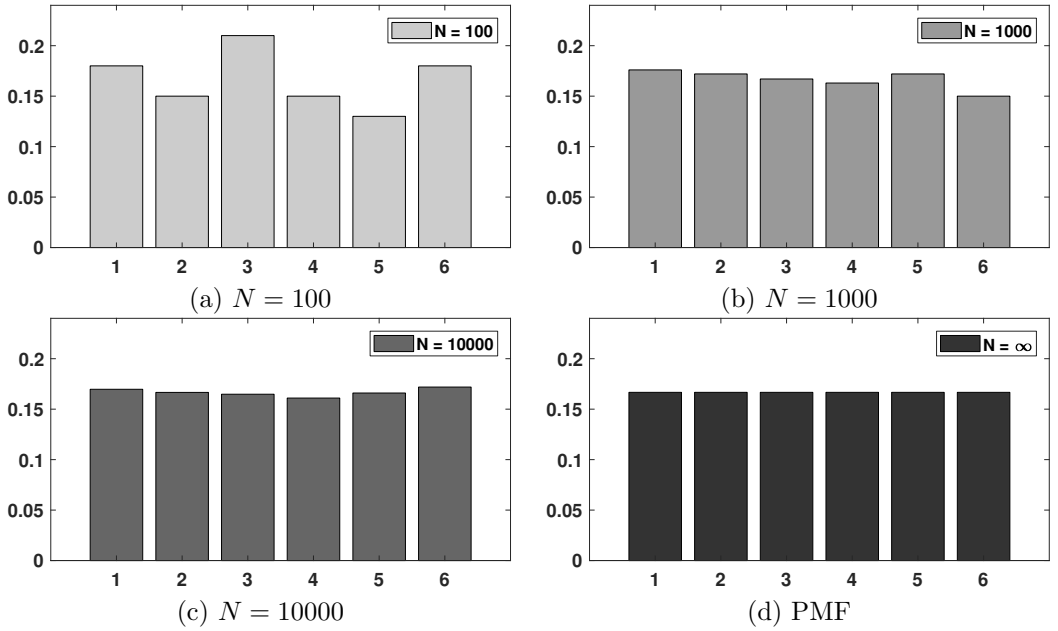


Figure 3.8: Histogram and PMF, when throwing a fair die N times. As N increases, the histograms are becoming more similar to the PMF.

histogram becomes more like the PMF. You can imagine that when N goes to infinity, the histogram will eventually become the PMF. Therefore, when given a dataset, one way to think of it is to treat the data as random realizations drawn from a certain PMF. The more data points you have, the closer the histogram will become to the PMF.

The MATLAB and Python codes used to generate **Figure 3.8** are shown below. The two commands we use here are `randi` (in MATLAB), which generates random integer numbers, and `hist`, which computes the heights and bin centers of a histogram. In Python, the corresponding commands are `np.random.randint` and `plt.hist`. Note that because of the different indexing schemes in MATLAB and Python, we offset the maximum index in `np.random.randint` to 7 instead of 6. Also, we shift the x-axes so that the bars are centered at the integers.

```
% MATLAB code to generate the histogram
x = [1 2 3 4 5 6];
q = randi(6,100,1);

figure;
[num,val] = hist(q,x-0.5);
bar(num/100,'FaceColor',[0.8, 0.8,0.8]);
axis([0 7 0 0.24]);
```

```
# Python code generate the histogram
import numpy as np
import matplotlib.pyplot as plt
q = np.random.randint(7,size=100)
```

```
plt.hist(q+0.5,bins=6)
```

This **generative** perspective is illustrated in **Figure 3.9**. We assume that the underlying latent random variable has some PMF that can be described by a few parameters, e.g., the mean and variance. Given the data points, if we can infer these parameters, we might retrieve the entire PMF (up to the uncertainty level intrinsic to the dataset). We refer to this inverse process as statistical inference.

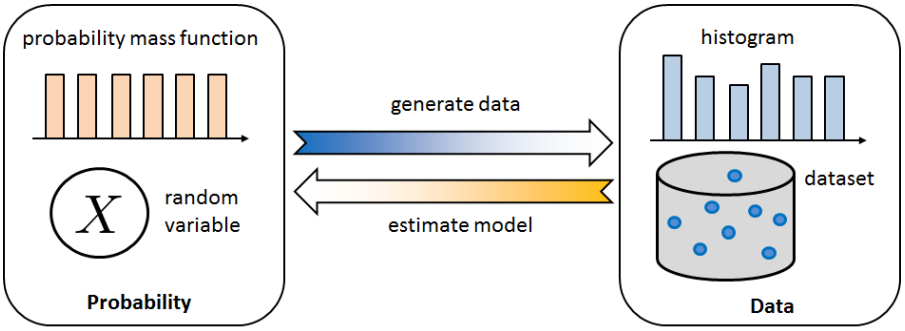


Figure 3.9: When analyzing a dataset, one can treat the data points are samples drawn according to a latent random variable with certain a PMF. The dataset we observe is often finite, and so the histogram we obtain is empirical. A major task in data analysis is statistical inference, which tries to retrieve the model information from the available measurements.

Returning to the question of why we need to understand the PMFs, the second part of the answer is the difference between **synthesis** and **analysis**. In synthesis, we start with a known random variable and generate samples according to the PMF underlying the random variable. For example, on a computer, we often start with a Gaussian random variable and generate random numbers according to the histogram specified by the Gaussian random variable. Synthesis is useful because we can predict what will happen. We can, for example, create millions of training samples to train a deep neural network. We can also evaluate algorithms used to estimate statistical quantities such as mean, variance, moments, etc., because the synthesis approach provides us with ground truth. In supervised learning scenarios, synthesis is vital to ensuring sufficient training data.

The other direction of synthesis is analysis. The goal is to start with a dataset and deduce the statistical properties of the dataset. For example, suppose we want to know whether the underlying model is indeed a Gaussian model. If we know that it is a Gaussian (or if we choose to use a Gaussian), we want to know the parameters that define this Gaussian. The analysis direction addresses this model selection and parameter estimation problem. Moving forward, once we know the model and the parameters, we can make a prediction or do recovery, both of which are ubiquitous in machine learning.

We summarize our discussions below, which is Key Concept 2 of this chapter.

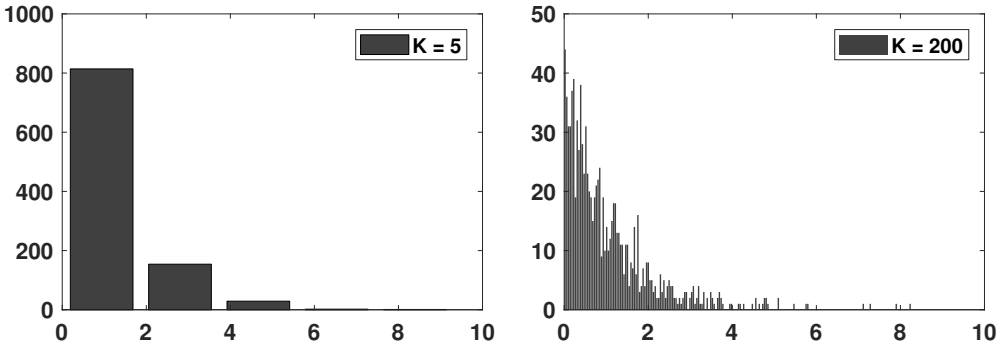
Key Concept 2: What are probability mass functions (PMFs)?

PMFs are the ideal histograms of random variables.

3.2.5 Estimating histograms from real data

The following discussions about histogram estimation can be skipped if it is your first time reading the book.

If you have a dataset, how would you plot the histogram? Certainly, if you have access to MATLAB or Python, you can call standard functions such as `hist` (in MATLAB) or `np.histogram` (in Python). However, when plotting a histogram, you need to specify the number of bins (or equivalently the width of bins). If you use larger bins, then you will have fewer bins with many elements in each bin. Conversely, if the bin width is too small, you may not have enough samples to fill the histogram. **Figure 3.10** illustrates two histograms in which the bins are respectively too large and too small.



(a) 5 bins (b) 200 bins

Figure 3.10: The width of the histogram has substantial influence on the information that can be extracted from the histogram.

The MATLAB and Python codes used to generate **Figure 3.10** are shown below. Note that here we are using an exponential random variable (to be discussed in Chapter 4). In MATLAB, calling an exponential random variable is done using `exprnd`, whereas in Python the command is `np.random.exponential`. For this experiment, we can specify the number of bins k, which can be set to $k = 200$ or $k = 5$. To suppress the Python output of the array, we can add a semicolon `;`. A final note is that `lambda` is a reserved variable in Python. Use something else.

```
% MATLAB code used to generate the plots
lambda = 1;
k      = 1000;
X      = exprnd(1/lambda,[k,1]);
[num,val] = hist(X,200);
bar(val,num,'FaceColor',[1, 0.5,0.5]);
```

```
# Python code used to generate the plots
import numpy as np
import matplotlib.pyplot as plt
lambd = 1
```

```
k      = 1000
X      = np.random.exponential(1/lambd, size=k)
plt.hist(X,bins=200);
```

In statistics, there are various rules to determine the bin width of a histogram. We mention a few of them here. Let K be the number of bins and N the number of samples.

- Square-root: $K = \sqrt{N}$
- Sturges' formula: $K = \log_2 N + 1$.
- Rice Rule: $K = 2\sqrt[3]{N}$
- Scott's normal reference rule: $K = \frac{\max X - \min X}{h}$, where $h = \frac{3.5\sqrt{\mathrm{Var}[X]}}{\sqrt[3]{N}}$ is the bin width.

For the example data shown in **Figure 3.10**, the histograms obtained using the above rules are given in **Figure 3.11**. As you can see, different rules have different suggested bin widths. Some are more conservative, e.g., using fewer bins, whereas some are less conservative. In any case, the suggested bin widths do seem to provide better histograms than the original ones in **Figure 3.10**. However, no bin width is the best for all purposes.

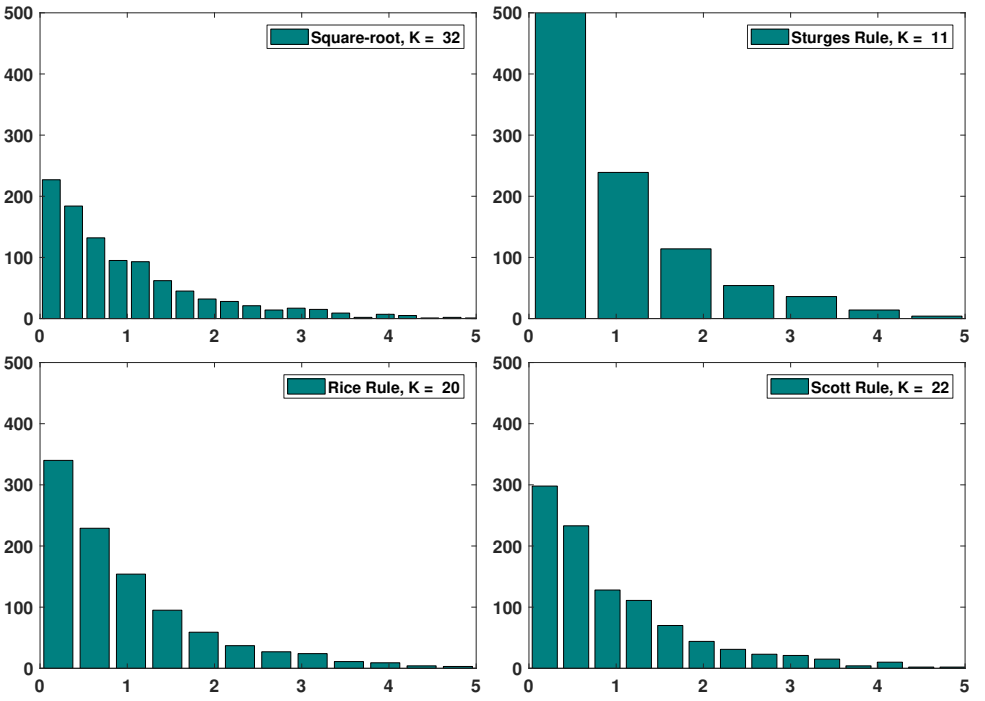


Figure 3.11: Histograms of a dataset using different bin width rules.

Beyond these predefined rules, there are also algorithmic tools to determine the bin width. One such tool is known as **cross-validation**. Cross-validation means defining some kind of **cross-validation score** that measures the statistical risk associated with the histogram. A histogram having a lower score has a lower risk, and thus it is a better histogram.

Note that the word "better" is relative to the optimality criteria associated with the cross-validation score. If you do not agree with our cross-validation score, our optimal bin width is not necessarily the one you want. In this case, you need to specify your optimality criteria.

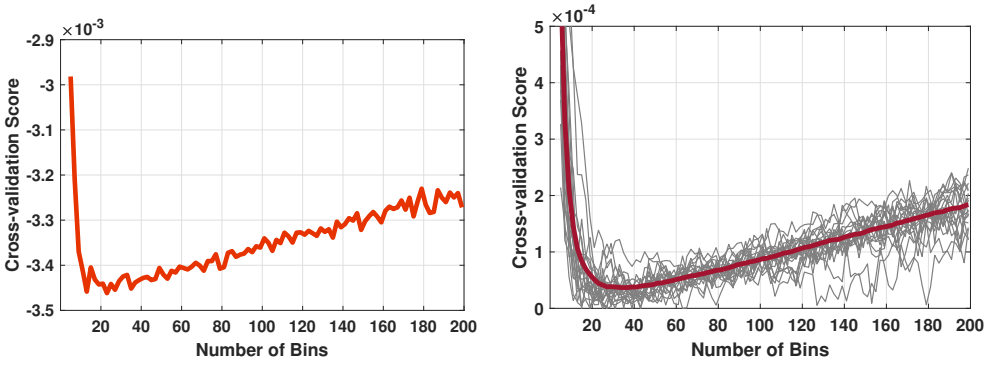
Theoretically, deriving a meaningful cross-validation score is beyond the scope of this book. However, it is still possible to understand the principle. Let h be the bin width of the histogram, K the number of bins, and N the number of samples. Given a dataset, we follow this procedure:

- Step 1: Choose a bin width h.
- Step 2: Construct a histogram from the data, using the bin width h. The histogram will have the empirical PMF values $\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_K$, which are the heights of the histograms normalized so that the sum is 1.
- Step 3: Compute the cross-validation score (see Wasserman, *All of Statistics*, Section 20.2):

$$J(h) = \frac{2}{(N-1)h} - \frac{N+1}{(N-1)h}\left(\widehat{p}_1^2 + \widehat{p}_2^2 + \cdots + \widehat{p}_K^2\right) \tag{3.4}$$

- Repeat Steps 1, 2, 3, until we find an h that minimizes $J(h)$.

Note that when we use a different h, the PMF values $\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_K$ will change, and the number of bins K will also change. Therefore, when changing h, we are changing not only the terms in $J(h)$ that explicitly contain h but also terms that are implicitly influenced.



(a) One dataset (b) Average of many datasets

Figure 3.12: Cross-validation score for the histogram. (a) The score of one particular dataset. (b) The scores for many different datasets generated by the same model.

For the dataset we showed in **Figure 3.10**, the cross-validation score $J(h)$ is shown in **Figure 3.12**. We can see that although the curve is noisy, there is indeed a reasonably clear minimum happening around $20 \leq K \leq 30$, which is consistent with some of the rules.

The MATLAB and Python codes we used to generate **Figure 3.12** are shown below. The key step is to implement Equation (3.4) inside a for-loop, where the loop goes through the range of bins we are interested in. To obtain the PMF values $\widehat{p}_1, \ldots, \widehat{p}_K$, we call `hist` in MATLAB and `np.histogram` in Python. The bin width `h` is the number of samples `n` divided by the number of bins `m`.

```
% MATLAB code to perform the cross validation
lambda = 1;
n = 1000;
X = exprnd(1/lambda,[n,1]);
m = 6:200;
J = zeros(1,195);
for i=1:195
    [num,binc] = hist(X,m(i));
    h = n/m(i);
    J(i) = 2/((n-1)*h)-((n+1)/((n-1)*h))*sum( (num/n).^2 );
end
plot(m,J,'LineWidth',4,'Color',[0.9,0.2,0.0]);
```

```
# Python code to perform the cross validation
import numpy as np
import matplotlib.pyplot as plt
lambd = 1
n     = 1000
X     = np.random.exponential(1/lambd, size=n)
m     = np.arange(5,200)
J     = np.zeros((195))
for i in range(0,195):
  hist,bins = np.histogram(X,bins=m[i])
  h = n/m[i]
  J[i] = 2/((n-1)*h)-((n+1)/((n-1)*h))*np.sum((hist/n)**2)
plt.plot(m,J);
```

In **Figure 3.12**(b), we show another set of curves from the same experiment. The difference here is that we assume access to the true generative model so that we can generate the many datasets of the same distribution. In this experiment we generated $T = 1000$ datasets. We compute the cross-validation score $J(h)$ for each of the datasets, yielding T score functions $J^{(1)}(h), \ldots, J^{(T)}(h)$. We subtract the minimum because different realizations have different offsets. Then we compute the average:

$$\overline{J}(h) = \frac{1}{T} \sum_{t=1}^{T} \left\{ J^{(t)}(h) - \min_{h} \left\{ J^{(t)}(h) \right\} \right\}. \tag{3.5}$$

This gives us a smooth red curve as shown in **Figure 3.12**(b). The minimum appears to be at $N = 25$. This is the optimal N, concerning the cross-validation score, on the average of all datasets.

All rules, including cross-validation, are based on optimizing for a certain objective. Your objective could be different from our objective, and so our optimum is not necessarily your optimum. Therefore, cross-validation may not be the best. It depends on your problem.

End of the discussion.

3.3 Cumulative Distribution Functions (Discrete)

While the probability mass function (PMF) provides a complete characterization of a discrete random variable, the PMFs themselves are technically not "functions" because the impulses in the histogram are essentially delta functions. More formally, a PMF $p_X(k)$ should actually be written as

$$p_X(x) = \sum_{k \in X(\Omega)} \underbrace{p_X(k)}_{\text{PMF values}} \cdot \underbrace{\delta(x-k)}_{\text{delta function}}.$$

This is a train of delta functions, where the height is specified by the probability mass $p_X(k)$. For example, a random variable with PMF values

$$p_X(0) = \frac{1}{4}, \;\; p_X(1) = \frac{1}{2}, \;\; p_X(2) = \frac{1}{4}$$

will be expressed as

$$p_X(x) = \frac{1}{4}\delta(x) + \frac{1}{2}\delta(x-1) + \frac{1}{4}\delta(x-2).$$

Since delta functions need to be integrated to generate values, the typical things we want to do, e.g., integration and differentiation, are not as straightforward in the sense of Riemann-Stieltjes.

The way to handle the unfriendliness of the delta functions is to consider mild modifications of the PMF. This notation of "cumulative" distribution functions will allow us to resolve the delta function problems. We will defer the technical details to the next chapter. For the time being, we will briefly introduce the idea to prepare you for the technical discussion later.

3.3.1 Definition of the cumulative distribution function

Definition 3.3. *Let X be a discrete random variable with $\Omega = \{x_1, x_2, \ldots\}$. The* **cumulative distribution function** *(CDF) of X is*

$$F_X(x_k) \stackrel{\text{def}}{=} \mathbb{P}[X \leq x_k] = \sum_{\ell=1}^{k} p_X(x_\ell). \tag{3.6}$$

If $\Omega = \{\ldots, -1, 0, 1, 2, \ldots\}$, then the CDF of X is

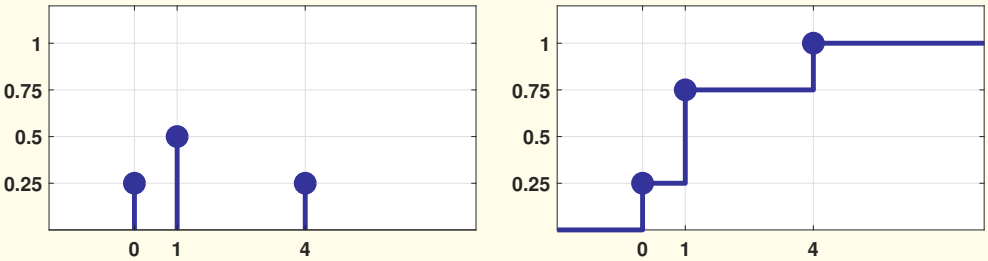
$$F_X(k) \stackrel{\text{def}}{=} \mathbb{P}[X \leq k] = \sum_{\ell=-\infty}^{k} p_X(\ell). \tag{3.7}$$

A CDF is essentially the cumulative sum of a PMF from $-\infty$ to x, where the variable x' in the sum is a dummy variable.

Example 3.6. Consider a random variable X with PMF $p_X(0) = \frac{1}{4}$, $p_X(1) = \frac{1}{2}$ and $p_X(4) = \frac{1}{4}$. The CDF of X can be computed as

$$F_X(0) = \mathbb{P}[X \leq 0] = p_X(0) = \frac{1}{4},$$
$$F_X(1) = \mathbb{P}[X \leq 1] = p_X(0) + p_X(1) = \frac{3}{4},$$
$$F_X(4) = \mathbb{P}[X \leq 4] = p_X(0) + p_X(1) + p_X(4) = 1.$$

As shown in **Figure 3.13**, the CDF of a discrete random variable is a staircase function.



(a) PMF $p_X(k)$ (b) CDF $F_X(k)$

Figure 3.13: Illustration of a PMF and a CDF.

The MATLAB code and the Python code used to generate **Figure 3.13** are shown below. The CDF is computed using the command `cumsum` in MATLAB and `np.cumsum` in Python.

```
% MATLAB code to generate a PMF and a CDF
p = [0.25 0.5 0.25];
x = [0 1 4];
F = cumsum(p);

figure(1);
stem(x,p,'.','LineWidth',4,'MarkerSize',50);
figure(2);
stairs([-4 x 10],[0 F 1],'.-','LineWidth',4,'MarkerSize',50);
```

```
% Python code to generate a PMF and a CDF
import numpy as np
import matplotlib.pyplot as plt
p = np.array([0.25, 0.5, 0.25])
x = np.array([0, 1, 4])
F = np.cumsum(p)

plt.stem(x,p,use_line_collection=True); plt.show()
plt.step(x,F); plt.show()
```

Why is CDF a better-defined function than PMF? There are technical reasons associated with whether a function is integrable. Without going into the details of these discussions, a short answer is that delta functions are defined through integrations; they are not functions. A delta function is defined as a function such that $\delta(x) = 0$ everywhere except at $x = 0$, and $\int_\Omega \delta(x)\, dx = 1$. On the other hand, a staircase function is always well-defined. The discontinuous points of a staircase can be well defined if we specify the gap between two consecutive steps. For example, in **Figure 3.13**, as soon as we specify the gap 1/4, 1/2, and 1/4, the staircase function is completely defined.

Example. **Figure 3.14** shows the empirical histogram of the English letters and the corresponding empirical CDF. We want to differentiate PMF versus histogram and CDF versus empirical CDF. The empirical CDF is the CDF computed from a finite dataset.

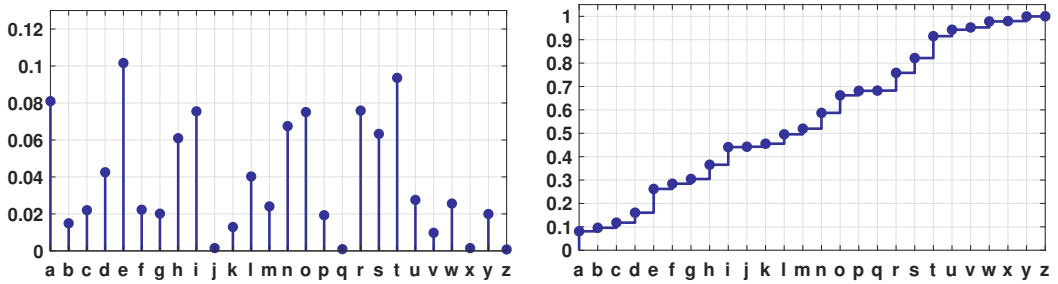


Figure 3.14: PMF and a CDF of the frequency of English letters.

3.3.2 Properties of the CDF

We observe from the example in Figure 3.13 that a CDF has several properties. First, being a staircase function, the CDF is non-decreasing. It can stay constant for a while, but it never drops. Second, the minimum value of a CDF is 0, whereas the maximum value is 1. It is 0 for any value that is smaller than the first state; it is 1 for any value that is larger than the last state. Third, the gap at each jump is exactly the probability mass at that state. Let us summarize these observations in the following theorem.

Theorem 3.2. *If X is a discrete random variable, then the CDF of X has the following properties:*

(i) The CDF is a sequence of **increasing** *unit steps.*

(ii) The **maximum** *of the CDF is when $x = \infty$: $F_X(+\infty) = 1$.*

(iii) The **minimum** *of the CDF is when $x = -\infty$: $F_X(-\infty) = 0$.*

(iv) The unit steps have **jumps** *at positions where $p_X(x) > 0$.*

Proof. Statement (i) can be seen from the summation

$$F_X(x) = \sum_{x' \le x} p_X(x').$$

Since the probability mass function is non-negative, the value of F_X is larger when the value of the argument is larger. That is, $x \leq y$ implies $F_X(x) \leq F_X(y)$. The second statement (ii) is true because the summation includes all possible states. So we have

$$F_X(+\infty) = \sum_{x'=-\infty}^{\infty} p_X(x') = 1.$$

Similarly, for the third statement (iii),

$$F_X(-\infty) = \sum_{x' \leq -\infty} p_X(x').$$

The summation is taken over an empty set, and so $F_X(-\infty) = 0$. Statement (iv) is true because the cumulative sum changes only when there is a non-zero mass in the PMF. □

As we can see in the proof, the basic argument of the CDF is the cumulative sum of the PMF. By definition, a cumulative sum always adds mass. This is why the CDF is always increasing, has 0 at $-\infty$, and has 1 at $+\infty$. This last statement deserves more attention. It implies that the unit step always has a **solid dot on the left**-hand side and an **empty dot on the right**-hand side, because when the CDF jumps, the final value is specified by the "$\leq$" sign in Equation (3.6). The technical term for this property is **right continuous**.

3.3.3 Converting between PMF and CDF

Theorem 3.3. *If X is a discrete random variable, then the PMF of X can be obtained from the CDF by*

$$p_X(x_k) = F_X(x_k) - F_X(x_{k-1}), \tag{3.8}$$

where we assumed that X has a countable set of states $\{x_1, x_2, \ldots\}$. If the sample space of the random variable X contains integers from $-\infty$ to $+\infty$, then the PMF can be defined as

$$p_X(k) = F_X(k) - F_X(k-1). \tag{3.9}$$

Example 3.7. Continuing with the example in **Figure 3.13**, if we are given the CDF

$$F_X(0) = \frac{1}{4}, \qquad F_X(1) = \frac{3}{4}, \qquad F_X(4) = 1,$$

how do we find the PMF? We know that the PMF will have non-negative values only at $x = 0, 1, 4$. For each of these x, we can show that

$$\begin{aligned} p_X(0) &= F_X(0) - F_X(-\infty) = \frac{1}{4} - 0 = \frac{1}{4}, \\ p_X(1) &= F_X(1) - F_X(0) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}, \\ p_X(4) &= F_X(4) - F_X(1) = 1 - \frac{3}{4} = \frac{1}{4}. \end{aligned}$$

3.4 Expectation

When analyzing data, it is often useful to extract certain key parameters such as the mean and the standard deviation. The mean and the standard deviation can be seen from the lens of random variables. In this section, we will formalize the idea using **expectation**.

3.4.1 Definition of expectation

Definition 3.4. *The* **expectation** *of a random variable X is*

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x\, p_X(x). \tag{3.10}$$

Expectation is the mean of the random variable X. Intuitively, we can think of $p_X(x)$ as the percentage of times that the random variable X attains the value x. When this percentage is multiplied by x, we obtain the contribution of each x. Summing over all possible values of x then yields the mean. To see this more clearly, we can write the definition as

$$\mathbb{E}[X] = \underbrace{\sum_{x \in X(\Omega)}}_{\text{sum over all states}} \underbrace{x}_{\text{a state } X \text{ takes}} \underbrace{p_X(x)}_{\text{the percentage}}.$$

Figure 3.15 illustrates a PMF that contains five states $x_1, \ldots, x_5$. Corresponding to each state are $p_X(x_1), \ldots, p_X(x_5)$. For this PMF to make sense, we must assume that $p_X(x_1) + \cdots + p_X(x_5) = 1$. To simplify notation, let us define $p_i \overset{\text{def}}{=} p_X(x_i)$. Then the expectation of X is just the sum of the products: value (x_i) times height (p_i). This gives $\mathbb{E}[X] = \sum_{i=1}^{5} x_i p_X(x_i)$.

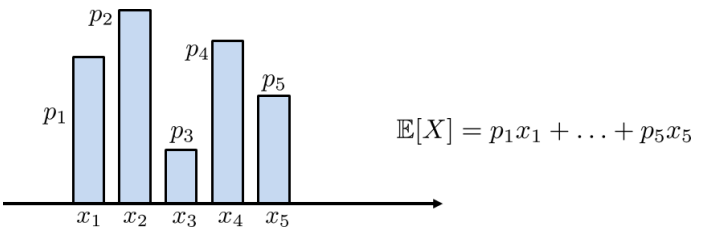


Figure 3.15: The expectation of a random variable is the sum of $x_i p_i$.

We emphasize that the definition of the expectation is exactly the same as the usual way we calculate the average of a dataset. When we calculate the average of a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$, we sum up these N samples and divide by the number of samples. This is what we called the empirical average or the sample average:

$$\text{average} = \frac{1}{N} \sum_{n=1}^{N} x^{(n)}. \tag{3.11}$$

Of course, in a typical dataset, these N samples often take distinct values. But suppose that among these N samples there are only K different values. For example, if we throw a die a million times, every sample we record will be one of the six numbers. This situation is illustrated in **Figure 3.16**, where we put the samples into the correct bin storing these values. In this case, to calculate the average we are effectively doing a binning:

$$\text{average} = \frac{1}{N}\sum_{k=1}^{K} \text{value } x_k \quad \times \quad \text{number of samples with value } x_k. \tag{3.12}$$

Equation (3.12) is *exactly* the same as Equation (3.11), as long as the samples can be grouped into K different values. With a little calculation, we can rewrite Equation (3.12) as

$$\text{average} = \underbrace{\sum_{k=1}^{K}}_{\text{sum of all states}} \underbrace{\text{value } x_k}_{\text{a state } X \text{ takes}} \times \underbrace{\frac{\text{number of samples with value } x_k}{N}}_{\text{the percentage}},$$

which is the same as the definition of expectation.

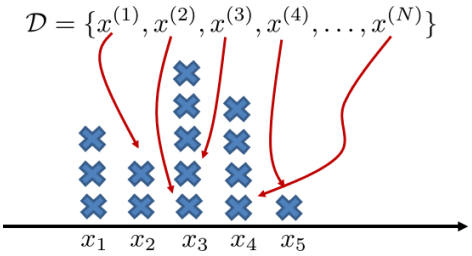


Figure 3.16: If we have a dataset $\mathcal{D}$ containing N samples, and if there are only K distinct values, we can effectively put these N samples into K bins. Thus, the "average" (which is the sum divided by the number N) is exactly the same as our definition of expectation.

The difference between $\mathbb{E}[X]$ and the average is that $\mathbb{E}[X]$ is computed from the *ideal* histogram, whereas average is computed from the *empirical* histogram. When the number of samples N approaches infinity, we expect the average to approximate $\mathbb{E}[X]$. However, when N is small, the empirical average will have random fluctuations around $\mathbb{E}[X]$. Every time we experiment, the empirical average may be slightly different. Therefore, we can regard $\mathbb{E}[X]$ as the *true average* of a certain random variable, and the empirical average as a *finite-sample average* based on the particular experiment we are working with. This summarizes Key Concept 3 of this chapter.

Key Concept 3: What is expectation?

Expectation = Mean = Average computed from a PMF.

If we are given a dataset on a computer, computing the mean can be done by calling the command `mean` in MATLAB and `np.mean` in Python. The example below shows the case of finding the mean of 10000 uniformly distributed random numbers.

```
% MATLAB code to compute the mean of a dataset
X  = rand(10000,1);
mX = mean(X);
```

```
# Python code to compute the mean of a dataset
import numpy as np
X  = np.random.rand(10000)
mX = np.mean(X)
```

Example 3.8. Let X be a random variable with PMF $p_X(0) = 1/4$, $p_X(1) = 1/2$ and $p_X(2) = 1/4$. We can show that the expectation is

$$\mathbb{E}[X] = (0)\underbrace{\left(\frac{1}{4}\right)}_{p_X(0)} + (1)\underbrace{\left(\frac{1}{2}\right)}_{p_X(1)} + (2)\underbrace{\left(\frac{1}{4}\right)}_{p_X(2)} = 1.$$

On MATLAB and Python, if we know the PMF then computing the expectation is straight-forward. Here is the code to compute the above example.

```
% MATLAB code to compute the expectation
p = [0.25 0.5 0.25];
x = [0 1 2];
EX = sum(p.*x);
```

```
# Python code to compute the expectation
import numpy as np
p = np.array([0.25, 0.5, 0.25])
x = np.array([0, 1, 2])
EX = np.sum(p*x)
```

Example 3.9. Flip an unfair coin, where the probability of getting a head is $\frac{3}{4}$. Let X be a random variable such that $X = 1$ means getting a head. Then we can show that $p_X(1) = \frac{3}{4}$ and $p_X(0) = \frac{1}{4}$. The expectation of X is therefore

$$\mathbb{E}[X] = (1)p_X(1) + (0)p_X(0) = (1)\left(\frac{3}{4}\right) + (0)\left(\frac{1}{4}\right) = \frac{3}{4}.$$

Center of mass. How would you interpret the result of this example? Does it mean that, on average, we will get 3/4 heads (but there is not anything called 3/4 heads!). Recall the definition of a random variable: it is a translator that translates a descriptive state to a number on the real line. Thus the expectation, which is an operation defined on the real line, can only tell us what is happening on the real line, not in the original sample

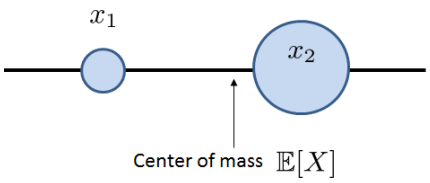


Figure 3.17: Center of mass. If a state x_2 is more influential than another state x_1, the center of mass $\mathbb{E}[X]$ will lean towards x_2.

space. On the real line, the expectation can be regarded as the **center of mass**, which is the point where the "forces" between the two states are "balanced". In **Figure 3.17** we depict a random variable with two states x_1 and x_2. The state x_1 has less influence (because $p_X(x_1)$ is smaller) than x_2. Therefore the center of mass is shifted towards x_2. This result shows us that the value $\mathbb{E}[X]$ is not necessarily in the sample space. $\mathbb{E}[X]$ is a deterministic number with nothing to do with the sample space.

Example 3.10. Let X be a random variable with PMF $p_X(k) = \frac{1}{2^k}$, for $k = 1, 2, 3, \ldots$. The expectation is

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k p_X(k) = \sum_{k=1}^{\infty} k \cdot \frac{1}{2^k}$$
$$= \frac{1}{2} \sum_{k=1}^{\infty} k \cdot \frac{1}{2^{k-1}} = \frac{1}{2} \cdot \frac{1}{(1-\frac{1}{2})^2} = 2.$$

On MATLAB and Python, if you want to verify this answer you can use the following code. Here, we approximate the infinite sum by a finite sum of $k = 1, \ldots, 100$.

```
% MATLAB code to compute the expectation
k = 1:100;
p = 0.5.^k;
EX = sum(p.*k);
```

```
# Python code to compute the expectation
import numpy as np
k = np.arange(100)
p = np.power(0.5,k)
EX = np.sum(p*k)
```

Example 3.11. Roll a die twice. Let X be the first roll and Y be the second roll. Let $Z = \max(X, Y)$. To compute the expectation $\mathbb{E}[Z]$, we first construct the sample space. Since there are two rolls, we can construct a table listing all possible pairs of outcomes. This will give us $\{(1,1), (1,2), \ldots, (6,6)\}$. Now, we calculate Z, which is the max of the two rolls. So if we have $(1,3)$, then the max will be 3, whereas if we have $(5,2)$, then the max will be 5. We can complete a table as shown below.

	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

This table tell us that Z has 6 states. The PMF of Z can be determined by counting the number of times a state shows up in the table. Thus, we can show that

$$p_Z(1) = \frac{1}{36},\ p_Z(2) = \frac{3}{36},\ p_Z(3) = \frac{5}{36},$$
$$p_Z(4) = \frac{7}{36},\ p_Z(5) = \frac{9}{36},\ p_Z(6) = \frac{11}{36}.$$

The expectation of Z is therefore

$$\begin{aligned}\mathbb{E}[Z] &= (1)\left(\frac{1}{36}\right) + (2)\left(\frac{3}{36}\right) + (3)\left(\frac{5}{36}\right) \\ &\quad + (4)\left(\frac{7}{36}\right) + (5)\left(\frac{9}{36}\right) + (6)\left(\frac{11}{36}\right) \\ &= \frac{161}{36}.\end{aligned}$$

Example 3.12. Consider a game in which we flip a coin 3 times. The reward of the game is

- \$1 if there are 2 heads
- \$8 if there are 3 heads
- \$0 if there are 0 or 1 head

There is a cost associated with the game. To enter the game, the player has to pay \$1.50. We want to compute the net gain, on average.

To answer this question, we first note that the sample space contains 8 elements: HHH, HHT, HTH, THH, THT, TTH, HTT, TTT. Let X be the number of heads. Then the PMF of X is

$$p_X(0) = \frac{1}{8},\ \ p_X(1) = \frac{3}{8},\ \ p_X(2) = \frac{3}{8},\ \ p_X(3) = \frac{1}{8}.$$

We then let Y be the reward. The PMF of Y can be found by "adding" the probabilities of X. This yields

$$p_Y(0) = p_X(0) + p_X(1) = \frac{4}{8},\ \ p_Y(1) = p_X(2) = \frac{3}{8},\ \ p_Y(8) = p_X(3) = \frac{1}{8}.$$

The expectation of Y is

$$\mathbb{E}[X] = (0)\left(\frac{4}{8}\right) + (1)\left(\frac{3}{8}\right) + (8)\left(\frac{1}{8}\right) = \frac{11}{8}.$$

Since the cost of the game is $\frac{12}{8}$, the net gain (on average) is $-\frac{1}{8}$.

3.4.2 Existence of expectation

Does every PMF have an expectation? No, because we can construct a PMF such that the expectation is undefined.

Example 3.13. Consider a random variable X with the following PMF:

$$p_X(k) = \frac{6}{\pi^2 k^2}, \qquad k = 1, 2, \ldots.$$

Using a result from algebra, one can show that $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$. Therefore, $p_X(k)$ is a legitimate PMF because $\sum_{k=1}^{\infty} p_X(k) = 1$. However, the expectation diverges, because

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=1}^{\infty} k p_X(k) \\ &= \frac{6}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{k} \to \infty,\end{aligned}$$

where the limit is due to the harmonic series[a]: $1 + \frac{1}{2} + \frac{1}{3} + \cdots = \infty$.

[a] https://en.wikipedia.org/wiki/Harmonic_series_(mathematics)

A PMF has an expectation when it is **absolutely summable**.

Definition 3.5. *A discrete random variable X is* **absolutely summable** *if*

$$\mathbb{E}[|X|] \stackrel{\text{def}}{=} \sum_{x \in X(\Omega)} |x|\, p_X(x) < \infty. \tag{3.13}$$

This definition tells us that not all random variables have a finite expectation. This is a very important mathematical result, but its practical implication is arguably limited. Most of the random variables we use in practice are absolutely summable. Also, note that the property of absolute summability applies to discrete random variables. For continuous random variables, we have a parallel concept called **absolute integrability**, which will be discussed in the next chapter.

3.4.3 Properties of expectation

The expectation of a random variable has several useful properties. We list them below. Note that these properties apply to both discrete and continuous random variables.

Theorem 3.4. *The expectation of a random variable X has the following properties:*

(i) **Function**. *For any function g,*

$$\mathbb{E}[g(X)] = \sum_{x \in X(\Omega)} g(x)\, p_X(x).$$

(ii) **Linearity**. *For any function g and h,*

$$\mathbb{E}[g(X) + h(X)] = \mathbb{E}[g(X)] + \mathbb{E}[h(X)].$$

(iii) **Scale**. *For any constant c,*

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

(iv) **DC Shift**. *For any constant c,*

$$\mathbb{E}[X + c] = \mathbb{E}[X] + c.$$

Proof of (i): A pictorial proof of (i) is shown in **Figure 3.18**. The key idea is a change of variable.

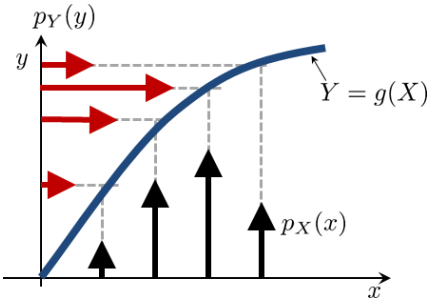


Figure 3.18: By letting $g(X) = Y$, the PMFs are not changed. What changes are the states.

When we have a function $Y = g(X)$, the PMF of Y will have impulses moved from x (the horizontal axis) to $g(x)$ (the vertical axis). The PMF values (i.e., the probabilities or the height of the stems), however, are not changed. If the mapping $g(X)$ is many-to-one, multiple PMF values will add to the same position. Therefore, when we compute $\mathbb{E}[g(X)]$, we compute the expectation along the vertical axis.

Practice Exercise 3.3. Prove statement (iii): For any constant c, $\mathbb{E}[cX] = c\mathbb{E}[X]$.

Solution. Recall the definition of expectation:

$$\mathbb{E}[cX] = \sum_{x \in X(\Omega)} (cx) p_X(x) = c \underbrace{\sum_{x \in X(\Omega)} x p_X(x)}_{=\mathbb{E}[X]} = c\mathbb{E}[X].$$

Statement (iii) is illustrated in **Figure 3.19**. Here, we assume that the original PMF has 3

states $X = 0, 1, 2$. We multiply X by a constant $c = 3$. This changes X to $cX = 0, 3, 6$. However, since the probabilities are not changed, the height of the PMF values remains. Therefore, when computing the expectation, we just multiply $\mathbb{E}[X]$ by c to get $c\mathbb{E}[X]$.

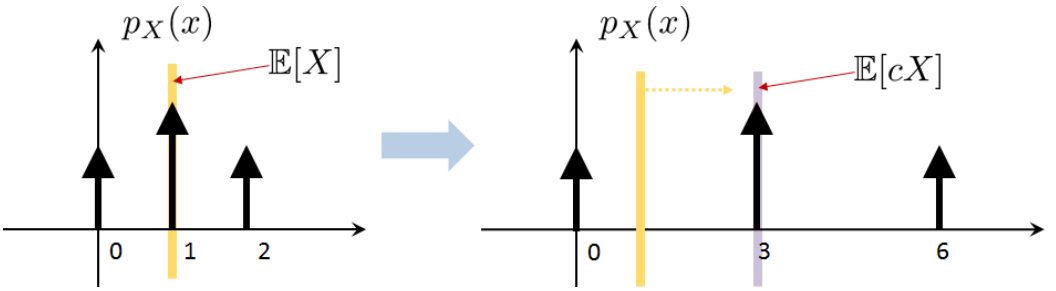


Figure 3.19: Pictorial representation of $\mathbb{E}[cX] = c\mathbb{E}[X]$. When we multiply X by c, we fix the probabilities but make the spacing between states wider/narrower.

Practice Exercise 3.4. Prove statement (ii): For any function g and h, $\mathbb{E}[g(X) + h(X)] = \mathbb{E}[g(X)] + \mathbb{E}[h(X)]$.

Solution. Recall the definition of expectation:

$$\begin{aligned}\mathbb{E}[g(X) + h(X)] &= \sum_{x \in X(\Omega)} [g(x) + h(x)] p_X(x) \\ &= \underbrace{\sum_{x \in X(\Omega)} g(x) p_X(x)}_{= \mathbb{E}[g(X)]} + \underbrace{\sum_{x \in X(\Omega)} h(x) p_X(x)}_{= \mathbb{E}[h(X)]} \\ &= \mathbb{E}[g(X)] + \mathbb{E}[h(X)].\end{aligned}$$

Practice Exercise 3.5. Prove statement (iv): For any constant c, $\mathbb{E}[X + c] = \mathbb{E}[X] + c$.

Solution. Recall the definition of expectation:

$$\begin{aligned}\mathbb{E}[X + c] &= \sum_{x \in X(\Omega)} (x + c) p_X(x) \\ &= \underbrace{\sum_{x \in X(\Omega)} x p_X(x)}_{= \mathbb{E}[X]} + c \cdot \underbrace{\sum_{x \in X(\Omega)} p_X(x)}_{= 1} \\ &= \mathbb{E}[X] + c.\end{aligned}$$

This result is illustrated in **Figure 3.20**. As we add a constant to the random variable, its PMF values remain the same but their positions are shifted. Therefore, when computing the mean, the mean will be shifted accordingly.

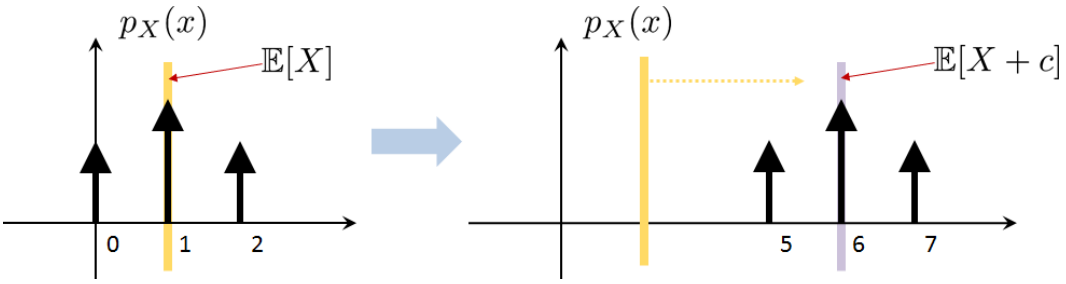


Figure 3.20: Pictorial representation of $\mathbb{E}[X+c] = \mathbb{E}[X]+c$. When we add c to X, we fix the probabilities and shift the entire PMF to the left or to the right.

Example 3.14. Let X be a random variable with four equally probable states $0, 1, 2, 3$. We want to compute the expectation $\mathbb{E}[\cos(\pi X/2)]$. To do so, we note that

$$\begin{aligned}\mathbb{E}[\cos(\pi X/2)] &= \sum_{x\in X(\Omega)} \cos\left(\frac{\pi X}{2}\right) p_X(x) \\ &= (\cos 0)\left(\frac{1}{4}\right) + (\cos\frac{\pi}{2})\left(\frac{1}{4}\right) + (\cos\frac{2\pi}{2})\left(\frac{1}{4}\right) + (\cos\frac{3\pi}{2})\left(\frac{1}{4}\right) \\ &= \frac{1+0+(-1)+0}{4} = 0.\end{aligned}$$

Example 3.15. Let X be a random variable with $\mathbb{E}[X] = 1$ and $\mathbb{E}[X^2] = 3$. We want to find the expectation $\mathbb{E}[(aX+b)^2]$. To do so, we realize that

$$\mathbb{E}[(aX+b)^2] \overset{(a)}{=} \mathbb{E}[a^2X^2 + 2abX + b^2] \overset{(b)}{=} a^2\mathbb{E}[X^2] + 2ab\mathbb{E}[X] + b^2 = 3a^2 + 2ab + b^2,$$

where (a) is due to expansion of the square, and (b) holds in two steps. The first step is to apply statement (ii) for individual functions of expectations, and the second step is to apply statement (iii) for scalar multiple of the expectations.

3.4.4 Moments and variance

Based on the concept of expectation, we can define a **moment**:

Definition 3.6. *The k**th moment** of a random variable X is*

$$\mathbb{E}[X^k] = \sum_x x^k\, p_X(x). \tag{3.14}$$

Essentially, the kth moment is the expectation applied to X^k. The definition follows from statement (i) of the expectation's properties. Using this definition, we note that $\mathbb{E}[X]$ is the first moment and $\mathbb{E}[X^2]$ is the second moment. Higher-order moments can be defined, but in practice they are less commonly used.

Example 3.16. Flip a coin 3 times. Let X be the number of heads. Then

$$p_X(0) = \frac{1}{8}, \quad p_X(1) = \frac{3}{8}, \quad p_X(2) = \frac{3}{8}, \quad p_X(3) = \frac{1}{8}.$$

The second moment $\mathbb{E}[X^2]$ is

$$\mathbb{E}[X^2] = (0)^2\left(\frac{1}{8}\right) + (1)^2\left(\frac{3}{8}\right) + (2)^2\left(\frac{3}{8}\right) + (4)^2\left(\frac{1}{8}\right) = 3.$$

Example 3.17. Consider a random variable X with PMF

$$p_X(k) = \frac{1}{2^k}, \qquad k = 1, 2, \ldots.$$

The second moment $\mathbb{E}[X^2]$ is

$$\begin{aligned}
\mathbb{E}[X^2] = \sum_{k=1}^{\infty} k^2 \left(\frac{1}{2}\right)^k &= \frac{1}{2^2}\sum_{k=1}^{\infty} k(k-1+1)\left(\frac{1}{2}\right)^{k-2} \\
&= \frac{1}{2^2}\sum_{k=1}^{\infty} k(k-1)\left(\frac{1}{2}\right)^{k-2} + \frac{1}{2^2}\sum_{k=1}^{\infty} k\left(\frac{1}{2}\right)^{k-2} \\
&= \frac{1}{2^2}\left(\frac{2}{(1-\frac{1}{2})^3}\right) + \frac{1}{2}\left(\frac{1}{(1-\frac{1}{2})^2}\right) = 6.
\end{aligned}$$

Using the second moment, we can define the **variance** of a random variable.

Definition 3.7. *The* **variance** *of a random variable X is*

$$\mathrm{Var}[X] = \mathbb{E}[(X-\mu)^2], \tag{3.15}$$

where $\mu = \mathbb{E}[X]$ is the expectation of X.

We denote σ^2 by $\mathrm{Var}[X]$. The square root of the variance, σ, is called the standard deviation of X. Like the expectation $\mathbb{E}[X]$, the variance $\mathrm{Var}[X]$ is computed using the ideal histogram PMF. It is the limiting object of the usual standard deviation we calculate from a dataset.

On a computer, computing the variance of a dataset is done by calling built-in commands such as `var` in MATLAB and `np.var` in Python. The standard deviation is computed using `std` and `np.std`, respectively.

```
% MATLAB code to compute the variance
X = rand(10000,1);
vX = var(X);
sX = std(X);
```

```
% Python code to compute the variance
import numpy as np
```

```
X = np.random.rand(10000)
vX = np.var(X)
sX = np.std(X)
```

What does the variance mean? It is a measure of the *deviation* of the random variable X relative to its mean. This deviation is quantified by the squared difference $(X-\mu)^2$. The expectation operator takes the average of the deviation, giving us a deterministic number $\mathbb{E}[(X-\mu)^2]$.

Theorem 3.5. *The variance of a random variable X has the following properties:*

(i) **Moment**.

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

(ii) **Scale**. *For any constant c,*

$$\mathrm{Var}[cX] = c^2\mathrm{Var}[X].$$

(iii) **DC Shift**. *For any constant c,*

$$\mathrm{Var}[X+c] = \mathrm{Var}[X].$$

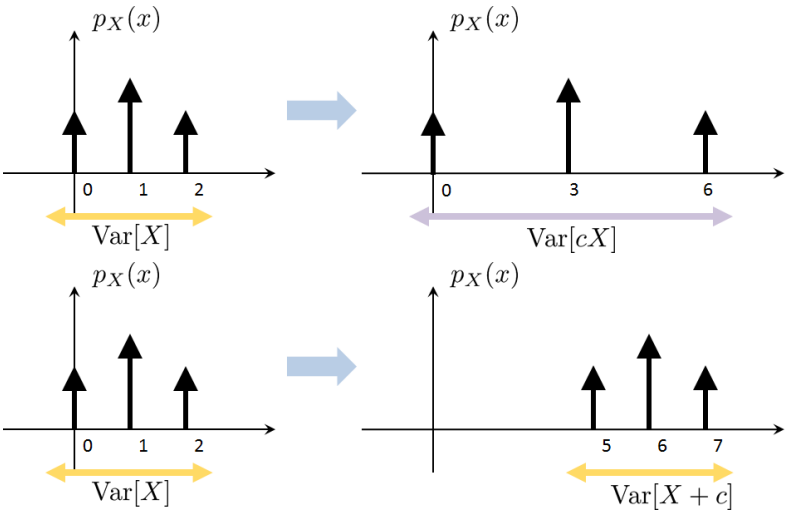


Figure 3.21: Pictorial representations of $\mathrm{Var}[cX] = c^2\mathrm{Var}[X]$ and $\mathrm{Var}[X+c] = \mathrm{Var}[X]$.

Practice Exercise 3.6. Prove Theorem 3.5 above.

Solution. For statement (i), we show that

$$\mathrm{Var}[X] = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2X\mu + \mu^2] = \mathbb{E}[X^2] - \mu^2.$$

Statement (ii) holds because $\mathbb{E}[cX] = c\mu$ and

$$\begin{aligned}\text{Var}[cX] &= \mathbb{E}[(cX - \mathbb{E}[cX])^2] \\ &= \mathbb{E}[(cX - c\mu)^2] = c^2\mathbb{E}[(X-\mu)^2] = c^2\text{Var}[X].\end{aligned}$$

Statement (iii) holds because

$$\text{Var}[X + c] = \mathbb{E}[((X + c) - \mathbb{E}[X + c])^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{Var}[X].$$

The properties above are useful in various ways. The first statement provides a link connecting variance and the second moment. Statement (ii) implies that when X is scaled by c, the variance should be scaled by c^2 because of the square in the second moment. Statement (iii) says that when X is shifted by a scalar c, the variance is unchanged. This is true because no matter how we shift the mean, the fluctuation of the random variable remains the same.

Practice Exercise 3.7. Flip a coin with probability p to get a head. Let X be a random variable denoting the outcome. The PMF of X is

$$p_X(0) = 1 - p, \qquad p_X(1) = p.$$

Find $\mathbb{E}[X]$, $\mathbb{E}[X^2]$ and $\text{Var}[X]$.

Solution. The expectation of X is

$$\mathbb{E}[X] = (0)p_X(0) + (1)p_X(1) = (0)(1-p) + (1)(p) = p.$$

The second moment is

$$\mathbb{E}[X^2] = (0)^2p_X(0) + (1)^2p_X(1) = p.$$

The variance is

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p).$$

3.5 Common Discrete Random Variables

In the previous sections, we have conveyed three key concepts: one about the random variable, one about the PMF, and one about the mean. The next step is to introduce a few commonly used discrete random variables so that you have something concrete in your "toolbox." As we have mentioned before, these predefined random variables should be studied from a **synthesis** perspective (sometimes called **generative**). The plan for this section is to introduce several models, derive their theoretical properties, and discuss examples.

Note that some extra effort will be required to understand the *origins* of the random variables. The origins of random variables are usually overlooked, but they are more important than the equations. For example, we will shortly discuss the Poisson random variable

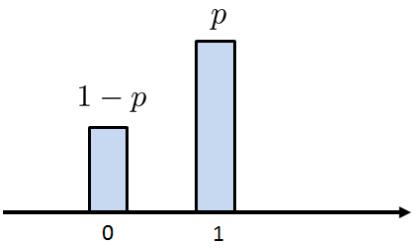


Figure 3.22: A Bernoulli random variable has two states with probability p and $1-p$.

and its PMF $p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}$. Why is the Poisson random variable defined in this way? If you know how the Poisson PMF was originally derived, you will understand the assumptions made during the derivation. Consequently, you will know why Poisson is a good model for internet traffic, recommendation scores, and image sensors for computer vision applications. You will also know under what situation the Poisson model will fail. Understanding the *physics* behind the probability models is the focus of this section.

3.5.1 Bernoulli random variable

We start discussing the simplest random variable, namely the **Bernoulli random variable**. A Bernoulli random variable is a *coin-flip* random variable. The random variable has two states: either 1 or 0. The probability of getting 1 is p, and the probability of getting 0 is $1-p$. See **Figure 3.22** for an illustration. Bernoulli random variables are useful for all kinds of binary state events: coin flip (H or T), binary bit (1 or 0), true or false, yes or no, present or absent, Democrat or Republican, etc.

To make these notions more precise, we define a Bernoulli random variable as follows.

Definition 3.8. *Let X be a* **Bernoulli random variable**. *Then, the PMF of X is*

$$p_X(0) = 1-p, \qquad p_X(1) = p,$$

where $0 < p < 1$ is called the Bernoulli parameter. We write

$$X \sim \text{Bernoulli}(p)$$

to say that X is drawn from a Bernoulli distribution with a parameter p.

In this definition, the parameter p controls the probability of obtaining 1. In a coin-flip event, p is usually $\frac{1}{2}$, meaning that the coin is fair. However, for biased coins p is not necessarily $\frac{1}{2}$. For other situations such as binary bits (0 or 1), the probability of obtaining 1 could be very different from the probability of obtaining 0.

In MATLAB and Python, generating Bernoulli random variables can be done by calling the binomial random number generator `np.random.binomial` (Python) and `binornd` (MATLAB). When the parameter `n` is equal to 1, the binomial random variable is equivalent to a Bernoulli random variable. The MATLAB and Python codes to synthesize a Bernoulli random variable are shown below.

```
% MATLAB code to generate 1000 Bernoulli random variables
p = 0.5;
n = 1;
X = binornd(n,p,[1000,1]);
[num, ~] = hist(X, 10);
bar(linspace(0,1,10), num,'FaceColor',[0.4, 0.4, 0.8]);
```

```
# Python code to generate 1000 Bernoulli random variables
import numpy as np
import matplotlib.pyplot as plt
p = 0.5
n = 1
X = np.random.binomial(n,p,size=1000)
plt.hist(X,bins='auto')
```

An alternative method in Python is to call `stats.bernoulli.rvs` to generate random Bernoulli numbers.

```
# Python code to call scipy.stats library
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
p = 0.5
X  = stats.bernoulli.rvs(p,size=1000)
plt.hist(X,bins='auto');
```

Properties of Bernoulli random variables

Let us now derive a few key statistical properties of a Bernoulli random variable.

Theorem 3.6. *If* $X \sim \text{Bernoulli}(p)$*, then*

$$\mathbb{E}[X] = p, \qquad \mathbb{E}[X^2] = p, \qquad \text{Var}[X] = p(1-p).$$

Proof. The expectation can be computed as

$$\mathbb{E}[X] = (1)p_X(1) + (0)p_X(0) = (1)(p) + (0)(1-p) = p.$$

The second moment is

$$\mathbb{E}[X^2] = (1^2)(p) + (0^2)(1-p) = p.$$

Therefore, the variance is

$$\text{Var}[X] = \mathbb{E}[X^2] - \mu^2 = p - p^2 = p(1-p).$$

□

A useful property of the Python code is that we can construct an object `rv`. Then we can call `rv`'s attributes to determine its mean, variance, etc.

```
# Python code to generate a Bernoulli rv object
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
p = 0.5
rv = stats.bernoulli(p)
mean, var = rv.stats(moments='mv')
print(mean, var)
```

In both MATLAB and Python, we can plot the PMF of a Bernoulli random variable, such as the one shown in **Figure 3.23**. To do this in MATLAB, we call the function `binopdf`, with the evaluation points specified by `x`.

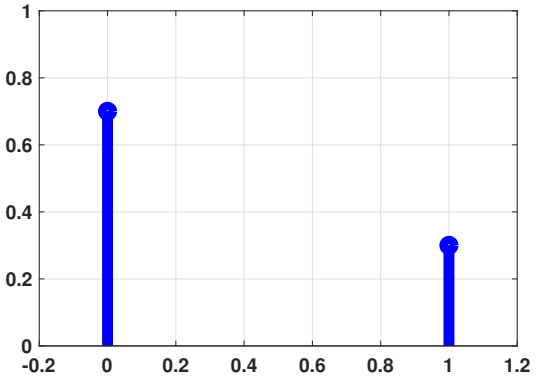


Figure 3.23: An example of a theoretical PMF (not the empirical histogram) plotted by MATLAB.

```
% MATLAB code to plot the PMF of a Bernoulli
p = 0.3;
x = [0,1];
f = binopdf(x,1,p);
stem(x, f, 'bo', 'LineWidth', 8);
```

In Python, we construct a random variable `rv`. With `rv`, we can call its PMF `rv.pmf`:

```
# Python code to plot the PMF of a Bernoulli
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
p = 0.3
rv = stats.bernoulli(p)
x  = np.linspace(0, 1, 2)
f  = rv.pmf(x)
plt.plot(x, f, 'bo', ms=10);
plt.vlines(x, 0, f, colors='b', lw=5, alpha=0.5);
```

When will a Bernoulli random variable have the maximum variance?

Let us take a look at the variance of the Bernoulli random variable. For any given p, the variance is $p(1-p)$. This is a quadratic equation. If we let $V(p) = p(1-p)$, we can show that the maximum is attained at $p = 1/2$. To see this, take the derivative of $V(p)$ with respect to p. This will give us $\frac{d}{dp}V(p) = 1 - 2p$. Equating to zero yields $1 - 2p = 0$, so $p = 1/2$. We know that $p = 1/2$ is a maximum and not a minimum point because the second order derivative $V''(p) = -2$, which is negative. Therefore $V(p)$ is maximized at $p = 1/2$. Now, since $0 \leq p \leq 1$, we also know that $V(0) = 0$ and $V(1) = 0$. Therefore, the variance is minimized at $p = 0$ and $p = 1$. **Figure 3.24** shows a graph of the variance.

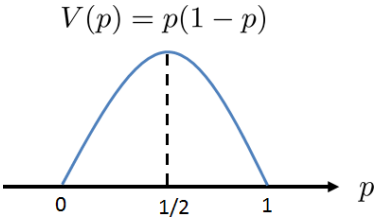


Figure 3.24: The variance of a Bernoulli reaches maximum at $p = 1/2$.

Does this result make sense? Why is the variance maximized at $p = 1/2$? If we think about this problem more carefully, we realize that a Bernoulli random variable represents a coin-flip experiment. If the coin is biased such that it always gives heads, on the one hand, it is certainly a bad coin. However, on the other hand, the variance is zero because there is nothing to vary; you will certainly get heads. The same situation happens if the coin is biased towards tails. However, if the coin is fair, i.e., $p = 1/2$, then the variance is large because we only have a 50% chance of getting a head or a tail whenever we flip a coin. Nothing is certain in this case. Therefore, the maximum variance happening at $p = 1/2$ matches our intuition.

Rademacher random variable

A slight variation of the Bernoulli random variable is the **Rademacher random variable**, which has two states: $+1$ and -1. The probability getting $+1$ and -1 is $1/2$. Therefore, the PMF of a Rademacher random variable is

$$p_X(-1) = \frac{1}{2}, \quad \text{and} \quad p_X(+1) = \frac{1}{2}.$$

Practice Exercise 3.8. Show that if X is a Rademacher random variable then $(X+1)/2 \sim \text{Bernoulli}(1/2)$. Also show the converse: If $Y \sim \text{Bernoulli}(1/2)$ then $2Y-1$ is a Rademacher random variable.

Solution. Since X can either be $+1$ or -1, we show that if $X = +1$ then $(X+1)/2 = 1$ and if $X = -1$ then $(X+1)/2 = 0$. The probabilities of getting $+1$ and -1 are equal. Thus, the probabilities of getting $(X+1)/2 = 1$ and 0 are also equal. So the resulting random variable is Bernoulli$(1/2)$. The other direction can be proved similarly.

Bernoulli in social networks: the Erdős-Rényi graph

The study of networks is a big branch of modern data science. It includes social networks, computer networks, traffic networks, etc. The history of network science is very long, but one of the most basic models of a network is the Erdős-Rényi graph, named after Paul Erdős and Alfréd Rényi. The underlying probabilistic model of the Erdős-Rényi graph is the Bernoulli random variable.

To see how a graph can be constructed from a Bernoulli random variable, we first introduce the concept of a **graph**. A graph contains two elements: nodes and edges. For node i and node j, we denote the edge connecting i and j as A_{ij}. Therefore, if we have N nodes, then we can construct a matrix $\boldsymbol{A}$ of size $N \times N$. We call this matrix the **adjacency matrix**. For example, the adjacency matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

will have edges for node pairs $(1, 2)$, $(1, 3)$, and $(3, 4)$. Note that in this example we assume that the adjacency matrix is symmetric, meaning that the graph is undirected. The "1" in the adjacency matrix indicates there is an edge, and "0" indicates there is no edge. So $\boldsymbol{A}$ represents a binary graph.

The Erdős-Rényi graph model says that the probability of getting an edge is an **independent** Bernoulli random variable. That is

$$A_{ij} \sim \text{Bernoulli}(p),$$

for $i < j$. If we model the graph in this way, then the parameter p will control the density of the graph. High values of p mean that there is a higher chance for an edge to be present.

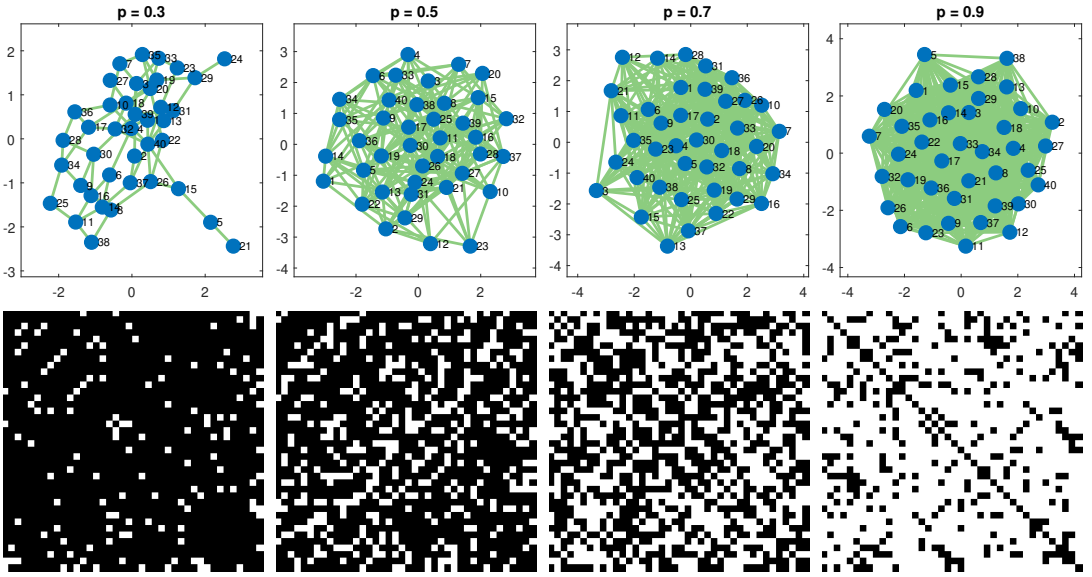


Figure 3.25: The Erdős-Rényi graph. [Top] The graphs. [Bottom] The adjacency matrices.

To illustrate the idea of an Erdős-Rényi graph, we show in **Figure 3.25** a graph of 40 nodes. The edges are randomly selected by flipping a Bernoulli random variable with parameter $p = 0.3, 0.5, 0.7, 0.9$. As we can see in the figure, a small value of p gives a graph with very sparse connectivity, whereas a large value of p gives a very densely connected graph. The bottom row of **Figure 3.25** shows the corresponding adjacency matrices. Here, a white pixel denotes "1" in the matrix and a black pixel denotes "0" in the matrix.

While Erdős-Rényi graphs are elementary, their variations can be realistic models of social networks. The **stochastic block model** is one such model. In a stochastic block model, nodes form small communities within a large network. For example, there are many majors in a university. Students within the same major tend to have more interactions than with students of another major. The stochastic block model achieves this goal by partitioning the nodes into communities. Within each community, the nodes can have a high degree of connectivity. Across different communities, the connectivity will be much lower. **Figure 3.26** illustrates a network and the corresponding adjacency matrix. In this example, the network has three communities.

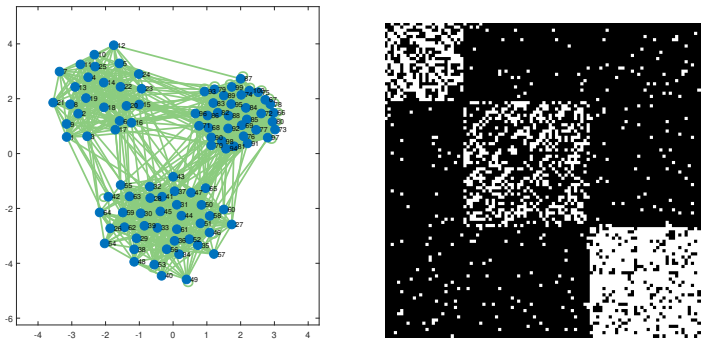


Figure 3.26: A stochastic block model containing three communities. [Left] The graph. [Right] The adjacency matrix.

In network analysis, one of the biggest problems is determining the community structure and recovering the underlying probabilities. The former task is about grouping the nodes into blocks. This is a nontrivial problem because in practice the nodes are never arranged nicely, as shown in **Figure 3.26**. For example, why should Alice be node 1 and Bob be node 2? Since we never know the correct ordering of the nodes, partitioning the nodes into blocks requires various estimation techniques such as clustering or iterative estimation. Recovering the underlying probability is also not easy. Given an adjacency matrix, why can we assume that the underlying network is a stochastic block model? Even if the model is correct, there will be imperfect grouping in the previous step. As such, estimating the underlying probability in the presence of these uncertainties would pose additional challenges.

Today, network analysis remains one of the hottest areas in data science. Its importance derives from its broad scope and impact. It can be used to analyze social networks, opinion polls, marketing, or even genome analysis. Nevertheless, the starting point of these advanced subjects is the Bernoulli random variable, the random variable of a coin flip!

3.5.2 Binomial random variable

Suppose we flip the coin n times count the number of heads. Since each coin flip is a random variable (Bernoulli), the sum is also a random variable. It turns out that this new random variable is the **binomial random variable**.

Definition 3.9. *Let X be a* **binomial random variable**. *Then, the PMF of X is*

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

where $0 < p < 1$ is the binomial parameter, and n is the total number of states. We write

$$X \sim \text{Binomial}(n, p)$$

to say that X is drawn from a binomial distribution with a parameter p of size n.

To understand the meaning of a binomial random variable, consider a simple experiment consisting of flipping a coin three times. We know that all possible cases are HHH, HHT, HTH, THH, TTH, THT, HTT and TTT. Now, suppose we define X = number of heads. We want to write down the probability mass function. Effectively, we ask: What is the probability of getting 0 head, one head, two heads, and three heads? We can, of course, count and get the answer right away for a fair coin. However, suppose the coin is unfair, i.e., the probability of getting a head is p whereas that of a tail is $1 - p$. The probability of getting each of the 8 cases is shown in **Figure 3.27** below.

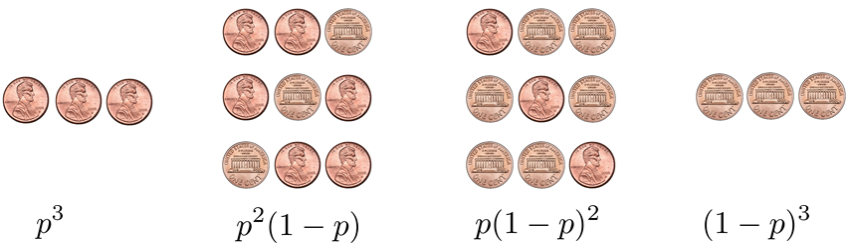


Figure 3.27: The probability of getting k heads out of $n = 3$ coins.

Here are the detailed calculations. Let us start with $X = 3$.

$$\begin{aligned} p_X(3) &= \mathbb{P}[\{\text{HHH}\}] \\ &= \mathbb{P}[\{\text{H}\} \cap \{\text{H}\} \cap \{\text{H}\}] \\ &\overset{(a)}{=} \mathbb{P}[\{\text{H}\}]\mathbb{P}[\{\text{H}\}]\mathbb{P}[\{\text{H}\}] \\ &\overset{(b)}{=} p^3, \end{aligned}$$

where (a) holds because the three events are independent. (Recall that if A and B are independent then $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$.) (b) holds because each $\mathbb{P}[\{\text{H}\}] = p$ by definition. With exactly the same argument, we can show that $p_X(0) = \mathbb{P}[\{\text{TTT}\}] = (1-p)^3$.

Now, let us look at $p_X(2)$, i.e., 2 heads. This probability can be calculated as follows:

$$\begin{aligned} p_X(2) &= \mathbb{P}[\{\text{HHT}\} \cup \{\text{HTH}\} \cup \{\text{THH}\}] \\ &\overset{(c)}{=} \mathbb{P}[\{\text{HHT}\}] + \mathbb{P}[\{\text{HTH}\}] + \mathbb{P}[\{\text{THH}\}] \\ &\overset{(d)}{=} p^2(1-p) + p^2(1-p) + p^2(1-p) = 3p^2(1-p), \end{aligned}$$

where (c) holds because the three events HHT, HTH and THH are disjoint in the sample space. Note that we are not using the independence argument in (c) but the disjoint argument. We should not confuse the two. The step in (d) uses independence, because each coin flip is independent.

The above calculation shows an interesting phenomenon: Although the three events HHT, HTH, and THH are different (in fact, disjoint), the number of heads in all the cases is the same. This happens because when counting the number of heads, the *ordering* of the heads and tails does not matter. So the same problem can be formulated as finding the number of combinations of { 2 heads and 1 tail }, which in our case is $\binom{3}{2} = 3$.

To complete the story, let us also try $p_X(1)$. This probability is

$$p_X(1) = \mathbb{P}[\{\text{TTH}\} \cup \{\text{HTT}\} \cup \{\text{THT}\}] = 3p(1-p)^2.$$

Again, we see that the combination $\binom{3}{1} = 3$ appears in front of the $p(1-p)^2$.

In general, the way to interpret the binomial random variable is to decouple the probabilities p, $(1-p)$, and the number of combinations $\binom{n}{k}$:

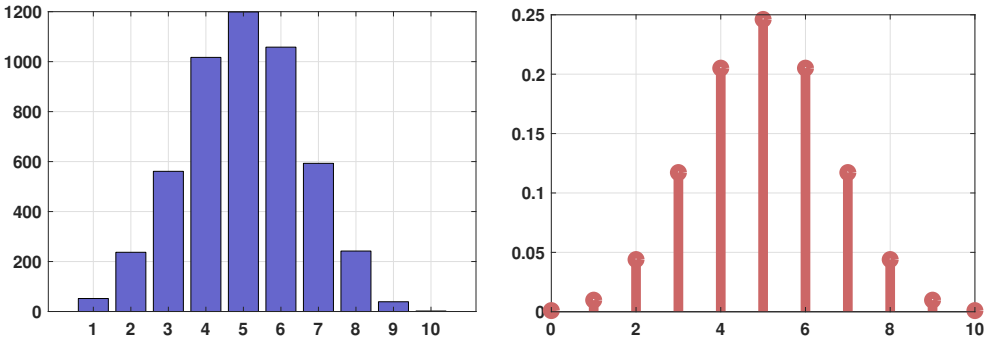
$$p_X(k) = \underbrace{\binom{n}{k}}_{\text{number of combinations}} \underbrace{p^k}_{\text{prob getting } k \text{ H's}} \underbrace{(1-p)^{n-k}}_{\text{prob getting } n-k \text{ T's}}.$$

The running index k should go with $0, 1, \ldots, n$. It starts with 0 because there could be zero heads in the sample space. Furthermore, we note that in this definition, two parameters are driving a binomial random variable: the number of Bernoulli trials n and the underlying probability for each coin flip p. As such, the notation for a binomial random variable is Binomial(n, p), with two arguments.

The histogram of a binomial random variable is shown in **Figure 3.28**(a). Here, we consider the example where $n = 10$ and $p = 0.5$. To generate the histogram, we use 5000 samples. In MATLAB and Python, generating binomial random variables as in **Figure 3.28**(a) can be done by calling `binornd` and `np.random.binomial`.

```
% MATLAB code to generate 5000 Binomial random variables
p = 0.5;
n = 10;
X = binornd(n,p,[5000,1]);
[num, ~] = hist(X, 10);
bar( num,'FaceColor',[0.4, 0.4, 0.8]);
```

```
# Python code to generate 5000 Binomial random variables
import numpy as np
import matplotlib.pyplot as plt
```



(a) Histogram based on 5000 samples (b) PMF

Figure 3.28: An example of a binomial distribution with $n = 10$, $p = 0.5$.

```
p = 0.5
n = 10
X = np.random.binomial(n,p,size=5000)
plt.hist(X,bins='auto');
```

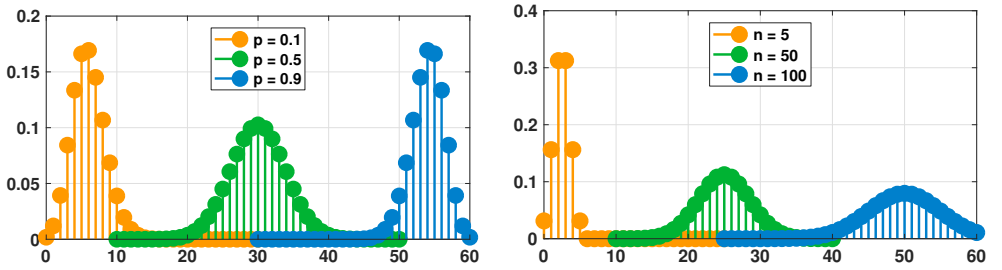
Generating the ideal PMF of a binomial random variable as shown in **Figure 3.28**(b) can be done by calling `binopdf` in MATLAB. In Python, we can define a random variable `rv` through `stats.binom`, and call the PMF using `rv.pmf`.

```
% MATLAB code to generate a binomial PMF
p = 0.5;
n = 10;
x = 0:10;
f = binopdf(x,n,p);
stem(x, f, 'o', 'LineWidth', 8, 'Color', [0.8, 0.4, 0.4]);
```

```
# Python code to generate a binomial PMF
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
p = 0.5
n = 10
rv = stats.binom(n,p)
x  = np.arange(11)
f  = rv.pmf(x)
plt.plot(x, f, 'bo', ms=10);
plt.vlines(x, 0, f, colors='b', lw=5, alpha=0.5);
```

The shape of the binomial PMF is shown in **Figure 3.29**. In this set of figures, we vary one of the two parameters n and p while keeping the other fixed. In **Figure 3.29**(a), we fix $n = 60$ and plot three sets of $p = 0.1, 0.5, 0.9$. For small p the PMF is skewed towards the left, and for large p the PMF is skewed toward the right. **Figure 3.29**(b) shows the PMF

for a fixed $p = 0.5$. As we increase n, the centroid of the PMF moves towards the right. Thus we should expect the mean of a binomial random variable to increase with p. Another interesting observation is that as n increases, the shape of the PMF approaches the Gaussian function (the bell-shaped curve). We will explain the reason for this when we discuss the Central Limit Theorem.



(a) $n = 60$ (b) $p = 0.5$

Figure 3.29: PMFs of a binomial random variable $X \sim \text{Binomial}(n, p)$. (a) We assume that $n = 60$. By varying the probability p, we see that the PMF shifts from the left to the right, and the shape changes. (b) We assume that $p = 0.5$. By varying the number of trials, the PMF shifts and the shape becomes more "bell-shaped."

The expectation, second moment, and variance of a binomial random variable are summarized in Theorem 3.7.

Theorem 3.7. *If* $X \sim \text{Binomial}(n, p)$, *then*

$$\begin{aligned} \mathbb{E}[X] &= np, \\ \mathbb{E}[X^2] &= np(np + (1-p)), \\ \text{Var}[X] &= np(1-p). \end{aligned}$$

We will prove that $\mathbb{E}[X] = np$ using the first principle. For $\mathbb{E}[X^2]$ and $\text{Var}[X]$, we will skip the proofs here and will introduce a "shortcut" later.

Proof. Let us start with the definition.

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \underbrace{0 \cdot \frac{n!}{0!(n-0)!} p^0 (1-p)^{n-0}}_{0} + \sum_{k=1}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k}. \end{aligned}$$

Note that we have shifted the index from $k = 0$ to $k = 1$. Now let us apply a trick:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k-1+1)!} p^k (1-p)^{n-k}.\end{aligned}$$

Using this trick, we can show that

$$\begin{aligned}&\sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k-1+1)!} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} \frac{n!}{(k-1)!((n-1)-(k-1))!} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} \frac{n(n-1)!}{(k-1)!((n-1)-(k-1))!} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} (1-p)^{n-k}\end{aligned}$$

With a simple substitution of $\ell = k - 1$, the above equation can be rewritten as

$$\begin{aligned}\mathbb{E}[X] &= np \cdot \sum_{\ell=0}^{n-1} \frac{(n-1)!}{\ell!((n-1)-\ell)!} p^{\ell} (1-p)^{n-1-\ell} \\ &= np \cdot \underbrace{\sum_{\ell=0}^{n-1} \binom{n-1}{k} p^{\ell} (1-p)^{n-1-\ell}}_{\text{summing PMF of Binomial}(n-1,p)} = np.\end{aligned}$$

□

In MATLAB, the mean and variance of a binomial random variable can be found by calling the command `binostat(n,p)` (MATLAB).

In Python, the command is `rv = stats.binom(n,p)` followed by calling `rv.stats`.

```
% MATLAB code to compute the mean and var of a binomial rv
p = 0.5;
n = 10;
[M,V] = binostat(n, p)
```

```
# Python code to compute the mean and var of a binomial rv
import scipy.stats as stats
p = 0.5
n = 10
rv = stats.binom(n,p)
M, V = rv.stats(moments='mv')
print(M, V)
```

An alternative view of the binomial random variable. As we discussed, the origin of a binomial random variable is the sum of a sequence of Bernoulli random variables. Because of this intrinsic definition, we can derive some useful results by exploiting this fact. To do so, let us define $I_1, \ldots, I_n$ as a sequence of Bernoulli random variables with $I_j \sim \text{Bernoulli}(p)$ for all $i = 1, \ldots, n$. Then the resulting variable

$$X = I_1 + I_2 + \cdots + I_n$$

is a binomial random variable of size n and parameter p. Using this definition, we can compute the expectation as follows:

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[I_1 + I_2 + \cdots + I_n] \\
&\overset{(a)}{=} \mathbb{E}[I_1] + \mathbb{E}[I_2] + \cdots + \mathbb{E}[I_n] \\
&= p + p + \cdots + p \\
&= np.
\end{aligned}$$

In this derivation, the step (a) depends on a useful fact about expectation (which we have not yet proved): For any two random variables X and Y, it holds that $\mathbb{E}[X+Y] = \mathbb{E}[X] + E[Y]$. Therefore, we can show that the expectation of X is np. This line of argument not only simplifies the proof but also provides a good intuition of the expectation. If each coin flip has an expectation of $\mathbb{E}[I_i] = p$, then the expectation of the sum should be simply n times of p, given np.

How about the variance? Again, we are going to use a very useful fact about variance: If two random variables X and Y are independent, then $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$. With this result, we can show that

$$\begin{aligned}
\text{Var}[X] &= \text{Var}[I_1 + \cdots + I_n] \\
&= \text{Var}[I_1] + \cdots + \text{Var}[I_n] \\
&= p(1-p) + \cdots + p(1-p) \\
&= np(1-p).
\end{aligned}$$

Finally, using the fact that $\text{Var}[X] = \mathbb{E}[X^2] - \mu^2$, we can show that

$$\begin{aligned}
\mathbb{E}[X^2] &= \text{Var}[X] + \mu^2 \\
&= np(1-p) + (np)^2.
\end{aligned}$$

Practice Exercise 3.9. Show that the binomial PMF sums to 1.

Solution. We use the binomial theorem to prove this result:

$$\sum_{k=0}^{n} p_X(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1.$$

The CDF of the binomial random variable is not very informative. It is basically the cumulative sum of the PMF:

$$F_X(k) = \sum_{\ell=0}^{k} \binom{n}{\ell} p^\ell (1-p)^{n-\ell}.$$

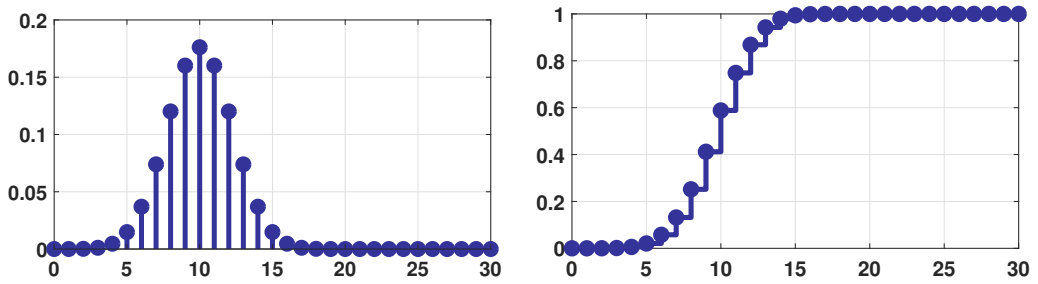


Figure 3.30: PMF and CDF of a binomial random variable $X \sim \text{Binomial}(n, p)$.

The shapes of the PMF and the CDF is shown in **Figure 3.30**.

In MATLAB, plotting the CDF of a binomial can be done by calling the function `binocdf`. You may also call `f = binopdf(x,n,p)`, and define `F = cumsum(f)` as the cumulative sum of the PMF. In Python, the corresponding command is `rv = stats.binom(n,p)` followed by `rv.cdf`.

```
% MATLAB code to compute the mean and var of a binomial rv
x = 0:10;
p = 0.5;
n = 10;
F = binocdf(x,n,p);
figure; stairs(x,F,'.-','LineWidth',4,'MarkerSize',30);
```

```
# Python code to compute the mean and var of a binomial rv
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
p = 0.5
n = 10
rv = stats.binom(n,p)
x  = np.arange(11)
F  = rv.cdf(x)
plt.plot(x, F, 'bo', ms=10);
plt.vlines(x, 0, F, colors='b', lw=5, alpha=0.5);
```

3.5.3 Geometric random variable

In some applications, we are interested in trying a binary experiment until we succeed. For example, we may want to keep calling someone until the person picks up the call. In this case, the random variable can be defined as the outcome of many failures followed by a final success. This is called the **geometric random variable**.

Definition 3.10. *Let X be a* **geometric random variable**. *Then, the PMF of X is*

$$p_X(k) = (1-p)^{k-1}p, \qquad k = 1, 2, \ldots,$$

where $0 < p < 1$ is the geometric parameter. We write

$$X \sim \text{Geometric}(p)$$

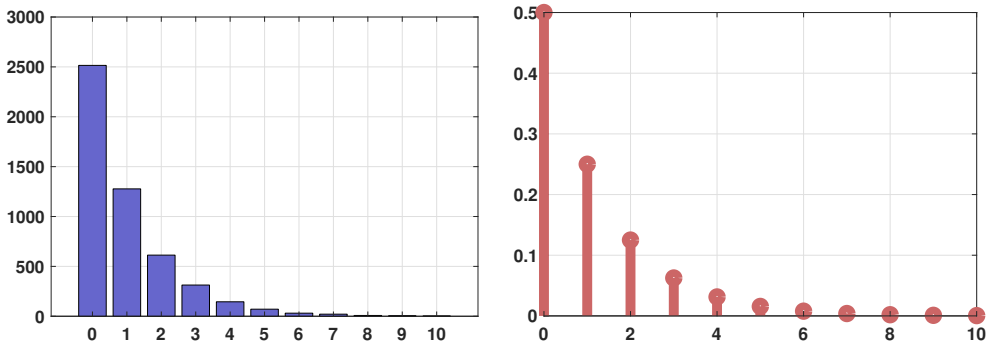
to say that X is drawn from a geometric distribution with a parameter p.

A geometric random variable is easy to understand. We define it as Bernoulli trials with $k-1$ consecutive failures followed by one success. This can be seen from the definition:

$$p_X(k) = \underbrace{(1-p)^{k-1}}_{k-1 \text{ failures}} \underbrace{p}_{\text{final success}}.$$

Note that in geometric random variables, there is no $\binom{n}{k}$ because we must have $k-1$ consecutive failures before one success. There is no alternative combination of the sequence.

The histogram and PMF of a geometric random variable are illustrated in **Figure 3.31**. Here, we assume that $p = 0.5$.



(a) Histogram based on 5000 samples (b) PMF

Figure 3.31: An example of a geometric distribution with $p = 0.5$.

In MATLAB, generating geometric random variables can be done by calling the commands `geornd`. In Python, it is `np.random.geometric`.

```
% MATLAB code to generate 1000 geometric random variables
p = 0.5;
X = geornd(p,[5000,1]);
[num, ~] = hist(X, 0:10);
bar(0:10, num, 'FaceColor',[0.4, 0.4, 0.8]);
```

```
# Python code to generate 1000 geometric random variables
import numpy as np
import matplotlib.pyplot as plt
p = 0.5
X = np.random.geometric(p,size=1000)
plt.hist(X,bins='auto');
```

To generate the PMF plots, in MATLAB we call `geopdf` and in Python we call `rv = stats.geom` followed by `rv.pmf`.

```
% MATLAB code to generate geometric PMF
p = 0.5; x = 0:10;
f = geopdf(x,p);
stem(x, f, 'o', 'LineWidth', 8, 'Color', [0.8, 0.4, 0.4]);
```

```
# Python code to generate 1000 geometric random variables
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
x  = np.arange(1,11)
rv = stats.geom(p)
f  = rv.pmf(x)
plt.plot(x, f, 'bo', ms=8, label='geom pmf')
plt.vlines(x, 0, f, colors='b', lw=5, alpha=0.5)
```

Practice Exercise 3.10. Show that the geometric PMF sums to one.

Solution. We can apply infinite series to show the result:

$$\begin{aligned}\sum_{k=1}^{\infty} p_X(k) &= \sum_{k=1}^{\infty}(1-p)^{k-1}p \\ &= p\cdot\sum_{k=1}^{\infty}(1-p)^{k-1}, \qquad \ell = k-1 \\ &= p\cdot\sum_{\ell=0}^{\infty}(1-p)^{\ell} \\ &= p\cdot\frac{1}{1-(1-p)} = 1.\end{aligned}$$

It is interesting to compare the shape of the PMFs for various values of p. In **Figure 3.32** we show the PMFs. We vary the parameter $p = 0.25, 0.5, 0.9$. For small p, the PMF starts with a low value and decays at a slow speed. The opposite happens for a large p, where the PMF starts with a high value and decays rapidly.

Furthermore, we can derive the following properties of the geometric random variable.

Theorem 3.8. *If* $X \sim \text{Geometric}(p)$, *then*

$$\mathbb{E}[X] = \frac{1}{p}, \qquad \mathbb{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}, \tag{3.16}$$
$$\text{Var}[X] = \frac{1-p}{p^2}.$$

Proof. We will prove that the mean is $1/p$ and leave the second moment and variance as

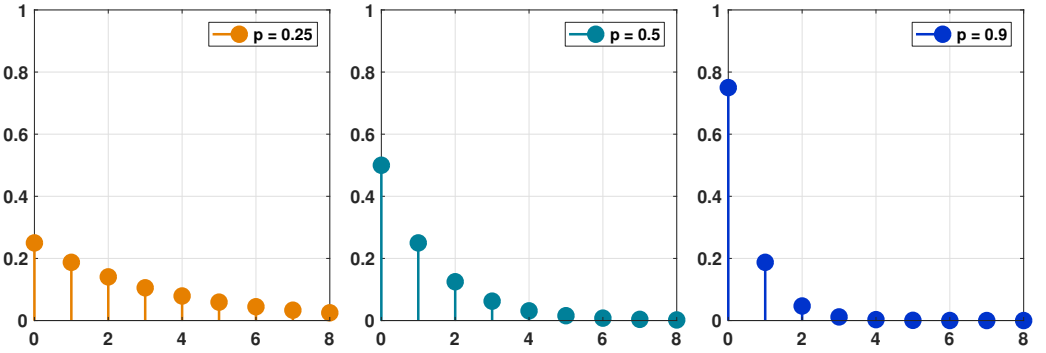


Figure 3.32: PMFs of a geometric random variable $X \sim \text{Geometric}(p)$.

an exercise.

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\left(\sum_{k=1}^{\infty} k(1-p)^{k-1}\right) \overset{(a)}{=} p\left(\frac{1}{(1-(1-p))^2}\right) = \frac{1}{p},$$

where (a) follows from the infinite series identity in Chapter 1.

□

3.5.4 Poisson random variable

In many physical systems, the arrivals of events are typically modeled as a Poisson random variable, e.g., photon arrivals, electron emissions, and telephone call arrivals. In social networks, the number of conversations per user can also be modeled as a Poisson. In e-commerce, the number of transactions per paying user is again modeled using a Poisson.

Definition 3.11. *Let X be a* **Poisson random variable**. *Then, the PMF of X is*

$$p_X(k) = \frac{\lambda^k}{k!}e^{-\lambda}, \qquad k = 0, 1, 2, \ldots,$$

where $\lambda > 0$ is the Poisson rate. We write

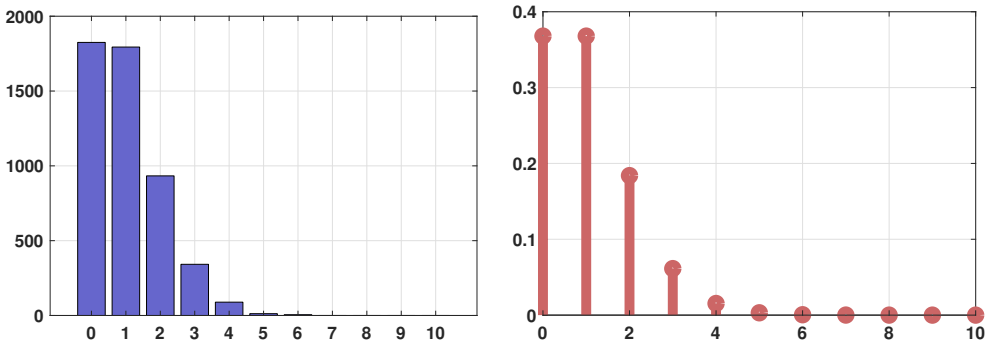
$$X \sim \text{Poisson}(\lambda)$$

to say that X is drawn from a Poisson distribution with a parameter λ.

In this definition, the parameter λ determines the **rate** of the arrival. The histogram and PMF of a Poisson random variable are illustrated in **Figure 3.33**. Here, we assume that $\lambda = 1$.

The MATLAB code and Python code used to generate the histogram are shown below.

```
% MATLAB code to generate 5000 Poisson numbers
lambda = 1;
X = poissrnd(lambda,[5000,1]);
```



(a) Histogram based on 5000 samples (b) PMF

Figure 3.33: An example of a Poisson distribution with $\lambda = 1$.

```
[num, ~] = hist(X, 0:10);
bar(0:10, num, 'FaceColor',[0.4, 0.4, 0.8]);
```

```
# Python code to generate 5000 Poisson random variables
import numpy as np
import matplotlib.pyplot as plt
lambd = 1
X = np.random.poisson(lambd,size=5000)
plt.hist(X,bins='auto');
```

For the PMF, in MATLAB we can call `poisspdf`, and in Python we can call `rv.pmf` with `rv = stats.poisson`.

```
% MATLAB code to plot the Poisson PMF
lambda = 1;
x = 0:10;
f = poisspdf(x,lambda);
stem(x, f, 'o', 'LineWidth', 8, 'Color', [0.8, 0.4, 0.4]);
```

```
# Python code to plot the Poisson PMF
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
x = np.arange(0,11)
rv = stats.poisson(lambd)
f  = rv.pmf(x)
plt.plot(x, f, 'bo', ms=8, label='geom pmf')
plt.vlines(x, 0, f, colors='b', lw=5, alpha=0.5)
```

The shape of the Poisson PMF changes with λ. As illustrated in **Figure 3.34**, $p_X(k)$ is more concentrated at lower values for smaller λ and becomes spread out for larger λ. Thus, we should expect that the mean and variance of a Poisson random variable will change

together as a function of λ. In the same figure, we show the CDF of a Poisson random variable. The CDF of a Poisson is

$$F_X(k) = \mathbb{P}[X \leq k] = \sum_{\ell=0}^{k} \frac{\lambda^\ell}{\ell!} e^{-\lambda}. \tag{3.17}$$

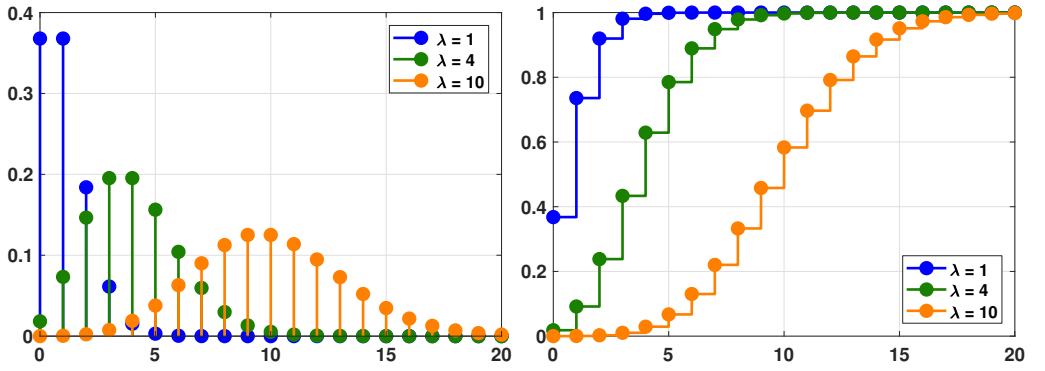


Figure 3.34: A Poisson random variable using different λ's. [Left] Probability mass function $p_X(k)$. [Right] Cumulative distribution function $F_X(k)$.

Example 3.18. Let X be a Poisson random variable with parameter λ. Find $\mathbb{P}[X > 4]$ and $\mathbb{P}[X \leq 5]$.

Solution.

$$\mathbb{P}[X > 4] = 1 - \mathbb{P}[X \leq 4] = 1 - \sum_{k=0}^{4} \frac{\lambda^k}{k!} e^{-\lambda},$$

$$\mathbb{P}[X \leq 5] = \sum_{k=0}^{5} \frac{\lambda^k}{k!} e^{-\lambda}.$$

Practice Exercise 3.11. Show that the Poisson PMF sums to 1.

Solution. We use the exponential series to prove this result:

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \cdot \underbrace{\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}}_{=e^{\lambda}} = 1.$$

Poisson random variables in practice

(1) Computational photography. In computational photography, the Poisson random variable is one of the most widely used models for photon arrivals. The reason pertains to the

origin of the Poisson random variable, which we will discuss shortly. When photons are emitted from the source, they travel through the medium as a sequence of independent events. During the integration period of the camera, the photons are accumulated to generate a voltage that is then translated to digital bits.

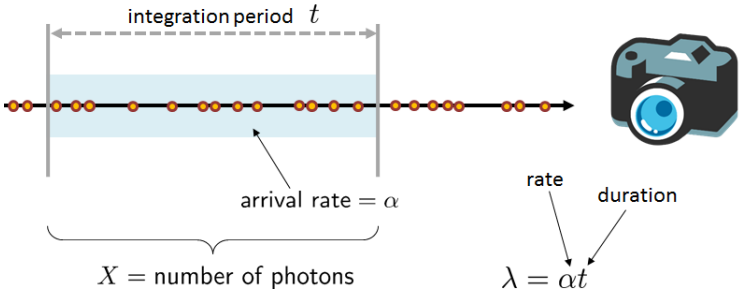


Figure 3.35: The Poisson random variable can be used to model photon arrivals.

If we assume that the photon arrival rate is α (photons per second), and suppose that the total amount of integration time is t, then the average number of photons that the sensor can see is αt. Let X be the number of photons seen during the integration time. Then if we follow the Poisson model, we can write down the PMF of X:

$$\mathbb{P}[X = k] = \frac{(\alpha t)^k}{k!} e^{-\alpha t}.$$

Therefore, if a pixel is bright, meaning that α is large, then X will have a higher likelihood of landing on a large number.

(2) Traffic model. The Poisson random variable can be used in many other problems. For example, we can use it to model the number of passengers on a bus or the number of spam phone calls. The required modification to **Figure 3.35** is almost trivial: merely replace the photons with your favorite cartoons, e.g., a person or a phone, as shown in **Figure 3.36**. In the United States, shared-ride services such as Uber and Lyft need to model the vacant cars and the passengers. As long as they have an arrival rate and certain degrees of independence between events, the Poisson random variable will be a good model.

As you can see from these examples, the Poisson random variable has broad applicability. Before we continue our discussion of its applications, let us introduce a few concepts related to the Poisson random variable.

Properties of a Poisson random variable

We now derive the mean and variance of a Poisson random variable.

Theorem 3.9. *If* $X \sim \text{Poisson}(\lambda)$, *then*

$$\mathbb{E}[X] = \lambda, \qquad \mathbb{E}[X^2] = \lambda + \lambda^2, \qquad \text{Var}[X] = \lambda. \tag{3.18}$$

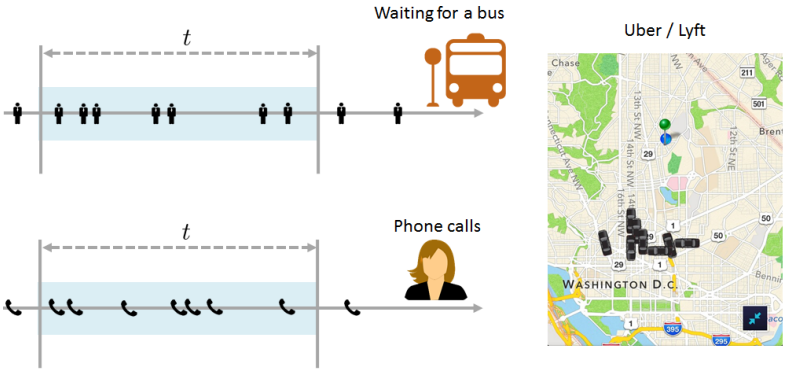


Figure 3.36: The Poisson random variable can be used to model passenger arrivals and the number of phone calls, and can be used by Uber or Lyft to provide shared rides.

Proof. Let us first prove the mean. It can be shown that

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda}$$
$$= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{\ell=0}^{\infty} \frac{\lambda^{\ell}}{\ell!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

The second moment can be computed as

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_{k=0}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\
&= \sum_{k=0}^{\infty} (k-1+1) \cdot \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\
&= \sum_{k=1}^{\infty} (k-1) \cdot \frac{\lambda^k}{(k-1)!} e^{-\lambda} + \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\
&= \lambda^2 \cdot \underbrace{\sum_{k=2}^{\infty} \frac{\lambda^{k-2} e^{-\lambda}}{(k-2)!}}_{=1} + \lambda \cdot \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!}}_{=1}.
\end{aligned}$$

The variance can be computed using $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mu^2$.

□

To compute the mean and variance of a Poisson random variable, we can call `poisstat` in MATLAB and `rv.stats(moments='mv')` in Python.

```
% MATLAB code to compute Poisson statistics
lambda = 1;
[M,V] = poisstat(lambda);
```

```
# Python code to compute Poisson statistics
import scipy.stats as stats
lambd = 1
rv = stats.poisson(lambd)
M, V = rv.stats(moments='mv')
```

The Poisson random variable is special in the sense that the mean and the variance are equal. That is, if the mean arrival number is higher, the variance is also higher. This is very different from some other random variables, e.g., the normal random variable where the mean and variance are independent. For certain engineering applications such as photography, this plays an important role in defining the signal-to-noise ratio. We will come back to this point later.

Origin of the Poisson random variable

We now address one of the most important questions about the Poisson random variable: Where does it come from? Answering this question is useful because the derivation process will reveal the underlying assumptions that lead to the Poisson PMF. When you change the problem setting, you will know when the Poisson PMF will hold and when the Poisson PMF will fail.

Our approach to addressing this problem is to consider the photon arrival process. (As we have shown, there is conceptually no difference if you replace the photons with pedestrians, passengers, or phone calls.) Our derivation follows the argument of J. Goodman, *Statistical Optics*, Section 3.7.2.

To begin with, we consider a photon arrival process. The total number of photons observed over an integration time t is defined as $X(t)$. Because $X(t)$ is a Poisson random variable, its arguments must be integers. The probability of observing $X(t) = k$ is therefore $\mathbb{P}[X(t) = k]$. **Figure 3.37** illustrates the notations and concepts.

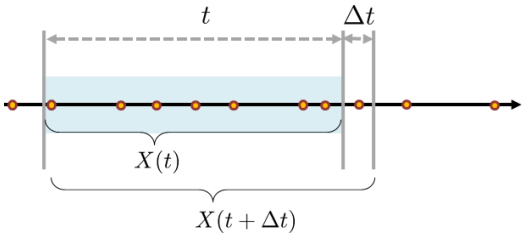


Figure 3.37: Notations for deriving the Poisson PMF.

We propose three hypotheses with the photon arrival process:

- For sufficiently small Δt, the probability of a small impulse occurring in the time interval $[t, t + \Delta t]$ is equal to the product of Δt and the rate λ, i.e.,

$$\mathbb{P}[X(t + \Delta t) - X(t) = 1] = \lambda \Delta t.$$

This is a linearity assumption, which typically holds for a short duration of time.

- For sufficiently small Δt, the probability that more than one impulse falls in Δt is negligible. Thus, we have that $\mathbb{P}[X(t+\Delta t) - X(t) = 0] = 1 - \lambda \Delta t$.
- The number of impulses in non-overlapping time intervals is independent.

The significance of these three hypotheses is that if the underlying photon arrival process violates any of these assumptions, then the Poisson PMF will not hold. One example is the presence of scattering effects, where a photon has a certain probability of going off due to the scattering medium and a certain probability of coming back. In this case, the events will no longer be independent.

Assuming that these hypotheses hold, then at time $t+\Delta t$, the probability of observing $X(t+\Delta t) = k$ can be computed as

$$\begin{aligned}
&\mathbb{P}[X(t+\Delta t) = k] \\
&= \mathbb{P}[X(t) = k] \cdot \underbrace{(1-\lambda\Delta t)}_{=\mathbb{P}[X(t+\Delta t)-X(t)=0]} + \mathbb{P}[X(t) = k-1] \cdot \underbrace{(\lambda \Delta t)}_{=\mathbb{P}[X(t+\Delta t)-X(t)=1]} \\
&= \mathbb{P}[X(t) = k] - \mathbb{P}[X(t) = k]\lambda\Delta t + \mathbb{P}[X(t) = k-1]\lambda\Delta t.
\end{aligned}$$

By rearranging the terms we show that

$$\frac{\mathbb{P}[X(t+\Delta t) = k] - \mathbb{P}[X(t) = k]}{\Delta t} = \lambda\Big(\mathbb{P}[X(t) = k-1] - \mathbb{P}[X(t) = k]\Big).$$

Setting the limit of $\Delta t \to 0$, we arrive at an ordinary differential equation

$$\frac{d}{dt}\mathbb{P}[X(t) = k] = \lambda\Big(\mathbb{P}[X(t) = k-1] - \mathbb{P}[X(t) = k]\Big). \tag{3.19}$$

We claim that the Poisson PMF, i.e.,

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

would solve this differential equation. To see this, we substitute the PMF into the equation. The left-hand side gives us

$$\begin{aligned}
\frac{d}{dt}\mathbb{P}[X(t) = k] &= \frac{d}{dt}\left(\frac{(\lambda t)^k}{k!} e^{-\lambda t}\right) \\
&= \lambda k \frac{(\lambda t)^{k-1}}{k!} e^{-\lambda t} + (-\lambda)\frac{(\lambda t)^k}{k!} e^{-\lambda t} \\
&= \lambda \frac{(\lambda t)^{k-1}}{k!} e^{-\lambda t} - \lambda \frac{(\lambda t)^k}{k!} e^{-\lambda t} \\
&= \lambda\Big(\mathbb{P}[X(t) = k-1] - \mathbb{P}[X(t) = k]\Big),
\end{aligned}$$

which is the right-hand side of the equation. To retrieve the basic form of Poisson, we can just set $t = 1$ in the PMF so that

$$\mathbb{P}[X(1) = k] = \frac{\lambda^k}{k!} e^{-\lambda}.$$

The origin of Poisson random variables

- We assume independent arrivals.
- Probability of seeing one event is linear with the arrival rate.
- Time interval is short enough so that you see either one event or no event.
- Poisson is derived by solving a differential equation based on these assumptions.
- Poisson becomes invalid when these assumptions are violated, e.g., in the case of scattering of photons due to turbid medium.

There is an alternative approach to deriving the Poisson PMF. The idea is to drive the parameter n in the binomial random variable to infinity while pushing p to zero. In this limit, the binomial PMF will converge to the Poisson PMF. We will discuss this shortly. However, we recommend the physics approach we have just described because it has a rich meaning and allows us to validate our assumptions.

Poisson approximation to binomial

We present one additional result about the Poisson random variable. The result shows that Poisson can be regarded as a limiting distribution of a binomial random variable.

Theorem 3.10. **(Poisson approximation to binomial)**. *For small p and large n,*

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda},$$

where $\lambda \stackrel{\text{def}}{=} np$.

Before we prove the result, let us see how close the approximation can be. In **Figure 3.38**, we show a binomial distribution and a Poisson approximation. The closeness of the approximation can easily be seen.

In MATLAB, the code to approximate a binomial distribution with a Poisson formula is shown below. Here, we draw 10,000 random binomial numbers and plot their histogram. On top of the plot, we use `poisspdf` to compute the Poisson PMF. This gives us **Figure 3.38**. A similar set of commands can be called in Python.

```
% MATLAB code to approximate binomial using Poisson
n = 1000; p = 0.05;
X = binornd(n,p,[10000,1]);
t = 0:100;
[num,val] = hist(X,t);
lambda = n*p;
f_pois = poisspdf(t,lambda);
bar(num/10000,'FaceColor',[0.9 0.9 0],'BarWidth',1); hold on;
plot(f_pois, 'LineWidth', 4);
```

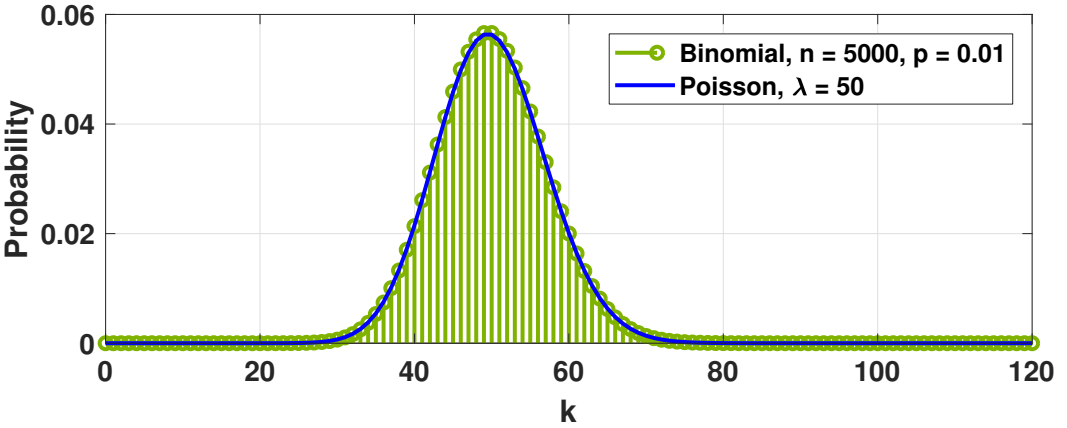


Figure 3.38: Poisson approximation of binomial distribution.

```
# Python code to approximate binomial using Poisson
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
n = 1000; p = 0.05
rv1 = stats.binom(n,p)
X   = rv1.rvs(size=10000)
plt.figure(1); plt.hist(X,bins=np.arange(0,100));
rv2 = stats.poisson(n*p)
f   = rv2.pmf(bin)
plt.figure(2); plt.plot(f);
```

Proof. Let $\lambda = np$. Then,

$$\begin{aligned}
\binom{n}{k} p^k (1-p)^{n-k} &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \frac{n(n-1)\cdots(n-k+1)}{n \cdot n \cdots n} \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \underbrace{(1)\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)}_{\to 1 \text{ as } n\to\infty} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1 \text{ as } n\to\infty} \left(1-\frac{\lambda}{n}\right)^{n} \\
&= \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^{n}.
\end{aligned}$$

We claim that $\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$. This can be proved by noting that

$$\log(1+x) \approx x, \qquad x \ll 1.$$

It then follows that $\log\left(1-\frac{\lambda}{n}\right) \approx -\frac{\lambda}{n}$. Hence, $\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}$

□

Example 3.19. Consider an optical communication system. The bit arrival rate is 10^9 bits/sec, and the probability of having one error bit is 10^{-9}. Suppose we want to find the probability of having five error bits in one second.

Let X be the number of error bits. In one second there are 10^9 bits. Since we do not know the location of these 5 bits, we have to enumerate all possibilities. This leads to a binomial distribution. Using the binomial distribution, we know that the probability of having k error bits is

$$\begin{aligned}\mathbb{P}[X=k] &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \binom{10^9}{k} (10^{-9})^k (1-10^{-9})^{10^9-k}.\end{aligned}$$

This quantity is difficult to calculate in floating-point arithmetic.

Using the Poisson to binomial approximation, we can see that the probability can be approximated by

$$\mathbb{P}[X=k] \approx \frac{\lambda^k}{k!} e^{-\lambda},$$

where $\lambda = np = 10^9(10^{-9}) = 1$. Setting $k=5$ yields $\mathbb{P}[X=5] \approx 0.003$.

Photon arrival statistics

Poisson random variables are useful in computer vision, but you may skip this discussion if it is your first reading of the book.

The strong connection between Poisson statistics and physics makes the Poisson random variable a very good fit for many physical experiments. Here we demonstrate an application in modeling photon shot noise.

An image sensor is a photon sensitive device which is used to detect incoming photons. In the simplest setting, we can model a pixel in the object plane as $X_{m,n}$, for some 2D coordinate $[m,n] \in \mathbb{R}^2$. Written as an array, an $M \times N$ image in the object plane can be visualized as

$$\boldsymbol{X} = \text{object} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,N} \\ \vdots & \vdots & \ddots & \vdots \\ X_{M,1} & X_{M,2} & \cdots & X_{M,N} \end{bmatrix}.$$

Without loss of generality, we assume that $X_{m,n}$ is normalized so that $0 \le X_{m,n} \le 1$ for every coordinate $[m,n]$. To model the brightness, we multiply $X_{m,n}$ by a scalar $\alpha > 0$. If a pixel $\alpha X_{m,n}$ has a large value, then it is a bright pixel; conversely, if $\alpha X_{m,n}$ has a small value, then it is a dark pixel. At a particular pixel location $[m,n] \in \mathbb{R}^2$, the observed pixel value $Y_{m,n}$ is a random variable following the Poisson statistics. This situation is illustrated

in **Figure 3.39**, where we see that an object-plane pixel will generate an observed pixel through the Poisson PMF.[1]

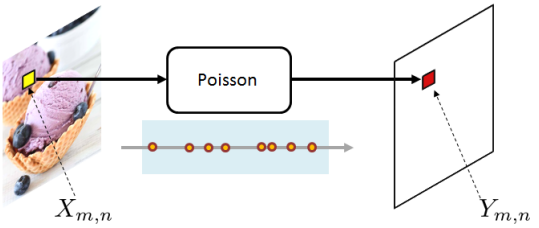


Figure 3.39: The image formation process is governed by the Poisson random variable. Given a pixel in the object plane $X_{m,n}$, the observed pixel $Y_{m,n}$ is a Poisson random variable with mean $\alpha X_{m,n}$. Therefore, a brighter pixel will have a higher Poisson mean, whereas a darker pixel will have a lower Poisson mean.

Written as an array, the image is

$$\begin{aligned}\boldsymbol{Y} = \text{observed image} &= \text{Poisson}\Big\{\alpha \boldsymbol{X}\Big\} \\ &= \begin{bmatrix} \text{Poisson}\{\alpha X_{1,1}\} & \text{Poisson}\{\alpha X_{1,2}\} & \cdots & \text{Poisson}\{\alpha X_{1,N}\} \\ \text{Poisson}\{\alpha X_{2,1}\} & \text{Poisson}\{\alpha X_{2,2}\} & \cdots & \text{Poisson}\{\alpha X_{2,N}\} \\ \vdots & \vdots & \ddots & \vdots \\ \text{Poisson}\{\alpha X_{M,1}\} & \text{Poisson}\{\alpha X_{M,2}\} & \cdots & \text{Poisson}\{\alpha X_{M,N}\} \end{bmatrix}.\end{aligned}$$

Here, by $\text{Poisson}\{\alpha X_{m,n}\}$ we mean that $Y_{m,n}$ is a random integer with probability mass

$$\mathbb{P}[Y_{m,n} = k] = \frac{[\alpha X_{m,n}]^k}{k!} e^{-\alpha X_{m,n}}.$$

Note that this model implies that the images seen by our cameras are more or less an array of Poisson random variables. (We say "more or less" because of other sources of uncertainties such as read noise, dark current, etc.) Because the observed pixels $Y_{m,n}$ are random variables, they fluctuate about the mean values, and hence they are noisy. We refer to this type of random fluctuation as the **shot noise**. The impact of the shot noise can be seen in **Figure 3.40**. Here, we vary the sensor gain level α. We see that for small α the image is dark and has much random fluctuation. As α increases, the image becomes brighter and the fluctuation becomes smaller.

In MATLAB, simulating the Poisson photon arrival process for an image requires the image-processing toolbox. The command to read an image is `imread`. Depending on the data type, the input array could be unit8 integers. To convert them to floating-point numbers between 0 and 1, we use the command `im2double`. Drawing Poisson measurements from the clean image is done using `poissrnd`. Finally, we can use `imshow` to display the image.

```
% MATLAB code to simulate a photon arrival process
x0 = im2double(imread('cameraman.tif'));
```

[1] The color of an image is often handled by a **color filter array**, which can be thought of as a wavelength selector that allows a specific wavelength to pass through.

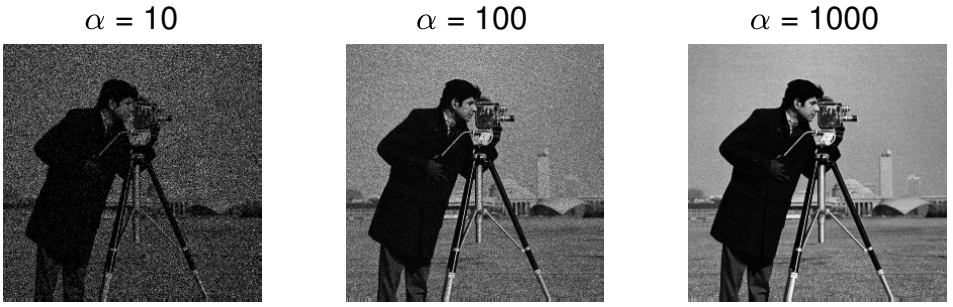


Figure 3.40: Illustration of the Poisson random variable in photographing images. Here, α denotes the gain level of the sensor: Larger α means that there are more photons coming to the sensor.

```
X  = poissrnd(10*x0);
figure(1); imshow(x0, []);
figure(2); imshow(X, []);
```

Similar commands can be found in Python with the help of the `cv2` library. When reading an image, we call `cv2.imread`. The option `0` is used to read a gray-scale image; otherwise, we will have a 3-channel color image. The division `/255` ensures that the input array ranges between 0 to 1. Generating the Poisson random numbers can be done using `np.random.poisson`, or by calling the statistics library with `stats.poisson.rvs(10*x0)`. To display the images, we call `plt.imshow`, with the color map option set to `cmap = 'gray'`.

```
# Python code code to simulate a photon arrival process
import numpy as np
import matplotlib.pyplot as plt
import cv2
x0 = cv2.imread('./cameraman.tif', 0)/255
plt.figure(1); plt.imshow(x0,cmap='gray');
X  = np.random.poisson(10*x0)
plt.figure(2); plt.imshow(X, cmap='gray');
```

Why study Poisson? What is shot noise?

- The Poisson random variable is used to model photon arrivals.
- Shot noise is the random fluctuation of the photon counts at the pixels. Shot noise is present even if you have an ideal sensor.

Signal-to-noise ratio of Poisson

Now let us answer a question we asked before. A Poisson random variable has a variance equal to the mean. Thus, if the scene is brighter, the variance will be larger. How come our simulation in **Figure 3.40** shows that the fluctuation becomes smaller as the scene becomes brighter?

The answer to this question lies in the **signal-to-noise ratio** (SNR) of the Poisson random variable. The SNR of an image defines its quality. The higher the SNR, the better the image. The mathematical definition of SNR is the ratio between the signal power and the noise power. In our case, the SNR is

$$\text{SNR} = \frac{\text{signal power}}{\text{noise power}} \overset{\text{def}}{=} \frac{\mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}} \overset{(a)}{=} \frac{\lambda}{\sqrt{\lambda}} = \sqrt{\lambda},$$

where $Y = Y_{m,n}$ is one of the observed pixels and $\lambda = \alpha X_{m,n}$ is the the corresponding object pixel. In this equation, the step (a) uses the properties of the Poisson random variable Y where $\mathbb{E}[Y] = \text{Var}[Y] = \lambda$. The result $\text{SNR} = \sqrt{\lambda}$ is very informative. It says that if the underlying mean photon flux (which is λ) increases, the SNR increases at a rate of $\sqrt{\lambda}$. So, yes, the variance becomes larger when the scene is brighter. However, the gain in signal $\mathbb{E}[Y]$ overrides the gain in noise $\sqrt{\text{Var}[Y]}$. As a result, the big fluctuation in bright images is compensated by the strong signal. Thus, to minimize the shot noise one has to use a longer exposure to increase the mean photon flux. When the scene is dark and the aperture is small, shot noise is unavoidable.

Poisson modeling is useful for describing the problem. However, the actual engineering question is that, given a noise observation $Y_{m,n}$, how would you reconstruct the clean image $X_{m,n}$? This is a very difficult **inverse problem**. The typical strategy is to exploit the spatial correlations between nearby pixels, e.g., usually smooth except along some sharp edges. Other information about the image, e.g., the likelihood of obtaining texture patterns, can also be leveraged. Modern image-processing methods are rich, ranging from classical filtering techniques to deep neural networks. Static images are easier to recover because we can often leverage multiple measurements of the same scene to boost the SNR. Dynamic scenes are substantially harder when we need to track the motion of any underlying objects. There are also newer image sensors with better photon sensitivity. The problem of imaging in the dark is an important research topic in **computational imaging**. New solutions are developed at the intersection of optics, signal processing, and machine learning.

The end of our discussions on photon statistics.

3.6 Summary

A **random variable** is so called because it can take more than one state. The probability mass function specifies the probability for it to land on a particular state. Therefore, whenever you think of a random variable you should immediately think of its PMF (or histogram if you prefer). The PMF is a unique characterization of a random variable. Two random variables with the same PMF are effectively the same random variables. (They are not identical because there could be measure-zero sets where the two differ.) Once you have the PMF, you can derive the CDF, expectation, moments, variance, and so on.

When your boss hands a dataset to you, which random variable (which model) should you use? This is a very practical and deep question. We highlight three steps for you to consider:

- (i) **Model selection**: Which random variable is the best fit for our problem? Sometimes we know by physics that, for example, photon arrivals or internet traffic follow a Poisson random variable. However, not all datasets can be easily described by simple models. The models we have learned in this chapter are called the **parametric** models because they are characterized by one or two parameters. Some datasets require nonparametric models, e.g., natural images, because they are just too complex. Some data scientists refer to deep neural networks as parametric models because the network weights are essentially the parameters. Some do not because when the number of parameters is on the order of millions, sometimes even more than the number of training samples, it seems more reasonable to call these models nonparametric. However, putting this debate aside, shortlisting a few candidate models based on prior knowledge is essential. Even if you use deep neural networks, selecting between convolutional structures versus long short-term memory models is still a legitimate task that requires an understanding of **your** problem.
- (ii) **Parameter estimation**: Suppose that you now have a candidate model; the next task is to estimate the model parameter using the available training data. For example, for Poisson we need to determine λ, and for binomial we need to determine (n, p). The estimation problem is an inverse problem. Often we need to use the PMF to construct certain optimization problems. By solving the optimization problem we will find the best parameter (for that particular candidate model). Modern machine learning is doing significantly better now than in the old days because optimization methods have advanced greatly.
- (iii) **Validation**. When each candidate model has been optimized to best fit the data, we still need to select the best model. This is done by running various testings. For example, we can construct a validation set and check which model gives us the best performance (such as classification rate or regression error). However, a model with the best validation score is not necessarily the best model. Your goal should be to seek a **good** model and not the **best** model because determining the best requires access to the testing data, which we do not have. Everything being equal, the common wisdom is to go with a simpler model because it is generally less susceptible to overfitting.

3.7 References

Probability textbooks

3-1 Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd Edition, 2008. Chapter 2.

3-2 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Chapter 3.

3-3 Athanasios Papoulis and S. Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th Edition, 2001. Chapters 3 and 4.

3-4 John A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, 2006. Chapters 2 and3.

3-5 Sheldon Ross, *A First Course in Probability*, Prentice Hall, 8th Edition, 2010. Chapter 4.

3-6 Henry Stark and John Woods, *Probability and Random Processes With Applications to Signal Processing*, Prentice Hall, 3rd Edition, 2001. Chapters 2 and 4.

Advanced probability textbooks

3-7 William Feller, *An Introduction to Probability Theory and Its Applications*, Wiley and Sons, 3rd Edition, 1950.

3-8 Andrey Kolmogorov, *Foundations of the Theory of Probability*, 2nd English Edition, Dover 2018. (Translated from Russian to English. Originally published in 1950 by Chelsea Publishing Company New York.)

Cross-validation

3-9 Larry Wasserman, *All of Statistics*, Springer 2004. Chapter 20.

3-10 Mats Rudemo, "Empirical Choice of Histograms and Kernel Density Estimators," *Scandinavian Journal of Statistics*, Vol. 9, No. 2 (1982), pp. 65-78.

3-11 David W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, 1992.

Poisson statistics

3-12 Joseph Goodman, *Statistical Optics*, Wiley, 2015. Chapter 3.

3-13 Henry Stark and John Woods, *Probability and Random Processes With Applications to Signal Processing*, Prentice Hall, 3rd edition, 2001. Section 1.10.

3.8 Problems

Exercise 1. (Video Solution)
Consider an information source that produces numbers k in the set $S_X = \{1, 2, 3, 4\}$. Find and plot the PMF in the following cases:

(a) $p_k = p_1/k$, for $k = 1, 2, 3, 4$. Hint: Find p_1.

(b) $p_{k+1} = p_k/2$ for $k = 1, 2, 3$.

(c) $p_{k+1} = p_k/2^k$ for $k = 1, 2, 3$.

(d) Can the random variables in parts (a)-(c) be extended to take on values in the set $\{1, 2, \ldots\}$? Why or why not? Hint: You may use the fact that the series $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ diverges.

Exercise 2. (Video Solution)
Two dice are tossed. Let X be the absolute difference in the number of dots facing up.

(a) Find and plot the PMF of X.

(b) Find the probability that $X \leq 2$.

(c) Find $\mathbb{E}[X]$ and $\text{Var}[X]$.

Exercise 3. (Video Solution)
Let X be a random variable with PMF $p_k = c/2^k$ for $k = 1, 2, \ldots$.

(a) Determine the value of c.

(b) Find $\mathbb{P}(X > 4)$ and $\mathbb{P}(6 \leq X \leq 8)$.

(c) Find $\mathbb{E}[X]$ and $\text{Var}[X]$.

Exercise 4.
Let X be a random variable with PMF $p_k = c/2^k$ for $k = -1, 0, 1, 2, 3, 4, 5$.

(a) Determine the value of c.

(b) Find $\mathbb{P}(1 \leq X < 3)$ and $\mathbb{P}(1 < X \leq 5)$.

(c) Find $\mathbb{P}[X^3 < 5]$.

(d) Find the PMF and the CDF of X.

Exercise 5. (Video Solution)
A modem transmits a $+2$ voltage signal into a channel. The channel adds to this signal a noise term that is drawn from the set $\{0, -1, -2, -3\}$ with respective probabilities $\{4/10, 3/10, 2/10, 1/10\}$.

(a) Find the PMF of the output Y of the channel.

(b) What is the probability that the channel's output is equal to the input of the channel?

(c) What is the probability that the channel's output is positive?

(d) Find the expected value and variance of Y.

Exercise 6.
On a given day, your golf score takes values from numbers 1 through 10, with equal probability of getting each one. Assume that you play golf for three days, and assume that your three performances are independent. Let X_1, X_2, and X_3 be the scores that you get, and let X be the minimum of these three numbers.

(a) Show that for any discrete random variable X, $p_X(k) = \mathbb{P}(X > k-1) - \mathbb{P}(X > k)$.

(b) What is the probability $\mathbb{P}(X_1 > k)$ for $k = 1, \ldots, 10$?

(c) Use (a), determine the PMF $p_X(k)$, for $k = 1, \ldots, 10$.

(d) What is the average score improvement if you play just for one day compared with playing for three days and taking the minimum?

Exercise 7. (Video Solution)
Let

$$g(X) = \begin{cases} 1, & \text{if } X > 10 \\ 0, & \text{otherwise.} \end{cases} \quad \text{and} \quad h(X) = \begin{cases} X - 10, & \text{if } X - 10 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find $\mathbb{E}[g(X)]$ for X as in Problem 1(a) with $S_X = \{1, \ldots, 15\}$.

(b) Find $\mathbb{E}[h(X)]$ for X as in Problem 1(b) with $S_X = \{1, \ldots, 15\}$.

Exercise 8. (Video Solution)
A voltage X is uniformly distributed in the set $\{-3, \ldots, 3, 4\}$.

(a) Find the mean and variance of X.

(b) Find the mean and variance of $Y = -2X^2 + 3$.

(c) Find the mean and variance of $W = \cos(\pi X/8)$.

(d) Find the mean and variance of $Z = \cos^2(\pi X/8)$.

Exercise 9. (Video Solution)

(a) If X is Poisson(λ), compute $\mathbb{E}[1/(X+1)]$.

(b) If X is Bernoulli(p) and Y is Bernoulli(q), compute $\mathbb{E}[(X+Y)^3]$ if X and Y are independent.

(c) Let X be a random variable with mean μ and variance σ^2. Let $\Delta(\theta) = \mathbb{E}[(X-\theta)^2]$. Find θ that minimizes the error $\Delta(\theta)$.

(d) Suppose that $X_1, \ldots, X_n$ are independent uniform random variables in $\{0, 1, \ldots, 100\}$. Evaluate $\mathbb{P}[\min(X_1, \ldots, X_n) > \ell]$ for any $\ell \in \{0, 1, \ldots, 100\}$.

Exercise 10. (Video Solution)

(a) Consider the binomial probability mass function $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$. Show that the mean is $\mathbb{E}[X] = np$.

(b) Consider the geometric probability mass function $p_X(k) = p(1-p)^k$ for $k = 0, 1, \ldots$. Show that the mean is $\mathbb{E}[X] = (1-p)/p$.

(c) Consider the Poisson probability mass function $p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$. Show that the variance is $\mathrm{Var}[X] = \lambda$.

(d) Consider the uniform probability mass function $p_X(k) = \frac{1}{L}$ for $k = 1, \ldots, L$. Show that the variance is $\mathrm{Var}[X] = \frac{L^2-1}{12}$. Hint: $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$ and $1^2 + 2^2 + \cdots + n^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$.

Exercise 11. (Video Solution)
An audio player uses a low-quality hard drive. The probability that the hard drive fails after being used for one month is 1/12. If it fails, the manufacturer offers a free-of-charge repair for the customer. For the cost of each repair, however, the manufacturer has to pay $20. The initial cost of building the player is $50, and the manufacturer offers a 1-year warranty. Within one year, the customer can ask for a free repair up to 12 times.

(a) Let X be the number of months when the player fails. What is the PMF of X? Hint: $\mathbb{P}[X = 1]$ may not be very high because if the hard drive fails it will be fixed by the manufacturer. Once fixed, the drive can fail again in the remaining months. So saying $X = 1$ is equivalent to saying that there is only one failure in the entire 12-month period.

(b) What is the average cost per player?

Exercise 12. (Video Solution)
A binary communication channel has a probability of bit error of $p = 10^{-6}$. Suppose that transmission occurs in blocks of 10,000 bits. Let N be the number of errors introduced by the channel in a transmission block.

(a) What is the PMF of N?

(b) Find $\mathbb{P}[N = 0]$ and $\mathbb{P}[N \leq 3]$.

(c) For what value of p will the probability of 1 or more errors in a block be 99%?

Hint: Use the Poisson approximation to binomial random variables.

Exercise 13. (Video Solution)
The number of orders waiting to be processed is given by a Poisson random variable with parameter $\alpha = \lambda/n\mu$, where λ is the average number of orders that arrive in a day, μ is the number of orders that an employee can process per day, and n is the number of employees. Let $\lambda = 5$ and $\mu = 1$. Find the number of employees required so the probability that more than four orders are waiting is less than 10%.

Hint: You need to use trial and error for a few n's.

Exercise 14.
Let X be the number of photons counted by a receiver in an optical communication system. It is known that X is a Poisson random variable with a rate λ_1 when a signal is present and a Poisson random variable with the rate $\lambda_0 < \lambda_1$ when a signal is absent. The probability that the signal is present is p. Suppose that we observe $X = k$ photons. We want to determine a threshold T such that if $k \geq T$ we claim that the signal is present, and if $k < T$ we claim that the signal is absent. What is the value of T?

Chapter 4

Continuous Random Variables

If you are coming to this chapter from Chapter 3, we invite you to take a 30-second pause and switch your mind from discrete events to continuous events. Everything is continuous now. The sample space is continuous, the event space is continuous, and the probability measure is continuous. Continuous random variables are similar in many ways to discrete random variables. They are characterized by the probability density functions (the continuous version of the probability mass functions); they have cumulative distribution functions; they have means, moments, and variances. The most significant difference is perhaps the use of integration instead of summation, but this change is conceptually straightforward, aside from the difficulties associated with integrating functions. So why do we need a separate chapter for continuous random variables? There are several reasons.

- First, how would you define the probability of a continuous event? Note that we cannot count because a continuous event is uncountable. There is also nothing called the probability mass because there are infinitely many masses. To define the probability of continuous events, we need to go back to our "slogan": **probability is a measure of the size of a set**. Because probability is a measure, we can speak meaningfully about the probability of continuous events so long as we have a well-defined measure for them. Defining such a measure requires some effort. We will develop the intuitions and the formal definitions in Section 4.1. In Section 4.2, we will discuss the expectation and variance of continuous random variables.
- The second challenge is the **unification** between continuous and discrete random variables. Since the two types of random variables ultimately measure the size of a set, it is natural to ask whether we can unify them. Our approach to unifying them is based on the cumulative distribution functions (CDFs), which are well-defined functions for discrete and continuous random variables. Based on the CDF and the fundamental theorem of calculus, we can show that the probability density functions and probability mass functions can be derived from the derivative of the CDFs. These will be discussed in Section 4.3, and in Section 4.4 we will discuss some additional results about the mode and median.
- The third challenge is to understand several widely used continuous random variables. We will discuss the uniform random variable and the exponential random variable in Section 4.5. Section 4.6 deals with the important topic of the **Gaussian random variable**. Where does a Gaussian random variable come from? Why does it have a bell

shape? Why are Gaussian random variables so popular in data science? What are the useful properties of Gaussian random variables? What are the relationships between a Gaussian random variable and other random variables? These important questions will be answered in Section 4.6.

- The final challenge is the **transformation** of random variables. Imagine that you have a random variable X and a function g. What will the probability mass/density function of $g(X)$ be? Addressing this problem is essential because almost all practical engineering problems involve the transformation of random variables. For example, suppose we have voltage measurements and we would like to compute the power. This requires taking the square of the voltage. We will discuss the transformation in Section 4.7, and we will also discuss an essential application in generating random numbers in Section 4.8.

4.1 Probability Density Function

4.1.1 Some intuitions about probability density functions

Let's begin by outlining some intuitive reasoning, which is needed to define the probability of continuous events properly. These intuitions are based on the fact that probability is a **measure**. In the following discussion you will see a sequence of logical arguments for constructing such a measure for continuous events. Some arguments are discussed in Chapter 2, but now we place them in the context of continuous random variables.

Suppose we are given an event A that is a subset in the sample space Ω, as illustrated in **Figure 4.1**. In order to calculate the probability of A, the measure perspective suggests that we consider the relative size of the set

$$\mathbb{P}[\{x \in A\}] = \frac{\text{“size” of } A}{\text{“size” of } \Omega}.$$

The right-hand side of this equation captures everything about the probability: It is a measure of the size of a set. It is relative to the sample space. It is a number between 0 and 1. It can be applied to discrete sets, and it can be applied to continuous sets.

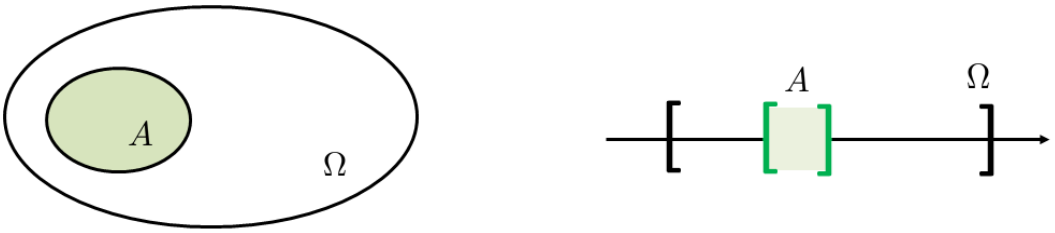


Figure 4.1: [Left] An event A in the sample space Ω. The probability that A happens can be calculated as the “size” of A relative to the “size” of Ω. [Right] A specific example on the real line. Note that the same definition of probability applies: The probability is the size of the interval A relative to that of the sample space Ω.

How do we measure the "size" of a continuous set? One possible way is by means of integrating the length, area, or volume covered by the set. Consider an example: Suppose that the sample space is the interval $\Omega = [0, 5]$ and the event is $A = [2, 3]$. To measure the "size" of A, we can integrate A to determine the length. That is,

$$\mathbb{P}[\{x \in [2,3]\}] = \frac{\text{"size" of } A}{\text{"size" of } \Omega} = \frac{\int_A \, dx}{\int_\Omega \, dx} = \frac{\int_2^3 \, dx}{\int_0^5 \, dx} = \frac{1}{5}.$$

Therefore, we have translated the "size" of a set to an integration. However, this definition is a very special case because when we calculate the "size" of a set, we treat all the elements in the set with equal importance. This is a strong assumption that will be relaxed later. But if you agree with this line of reasoning, we can rewrite the probability as

$$\begin{aligned}\mathbb{P}[\{x \in A\}] = \frac{\int_A \, dx}{\int_\Omega \, dx} &= \frac{\int_A \, dx}{|\Omega|} \\ &= \int_A \underbrace{\frac{1}{|\Omega|}}_{\text{equally important over } \Omega} dx.\end{aligned}$$

This equation says that under our assumption (that all elements are equiprobable), the probability of A is calculated as the integration of A using an integrand $1/|\Omega|$ (note that $1/|\Omega|$ is a constant with respect to x). If we evaluate the probability of another event B, all we need to do is to replace A with B and compute $\int_B \frac{1}{|\Omega|} \, dx$.

What happens if we want to relax the "equiprobable" assumption? Perhaps we can adopt something similar to the probability mass function (PMF). Recall that a PMF p_X evaluated at a point x is the probability that the state x happens, i.e., $p_X(x) = \mathbb{P}[X = x]$. So, $p_X(x)$ is the relative frequency of x. Following the same line of thinking, we can define a function f_X such that $f_X(x)$ tells us something related to the "relative frequency". To this end, we can treat f_X as a continuous histogram with infinitesimal bin width as shown in **Figure 4.2**. Using this f_X, we can replace the constant function $1/|\Omega|$ with the new function $f_X(x)$. This will give us

$$\mathbb{P}[\{x \in A\}] = \int_A \underbrace{f_X(x)}_{\text{replace } 1/|\Omega|} dx. \tag{4.1}$$

If we compare it with a PMF, we note that when X is discrete,

$$\mathbb{P}[\{x \in A\}] = \sum_{x \in A} p_X(x).$$

Hence, f_X can be considered a continuous version of p_X, although we do not recommend this way of thinking for the following reason: $p_X(x)$ is a legitimate probability, but $f_X(x)$ is not a probability. Rather, f_X is the **probability per unit length**, meaning that we need to integrate f_X (times dx) in order to generate a probability value. If we only look at f_X at a point x, then this point is a measure-zero set because the length of this set is zero.

Equation (4.1) should be familiar to you from Chapter 2. The function $f_X(x)$ is precisely the weighting function we described in that chapter.

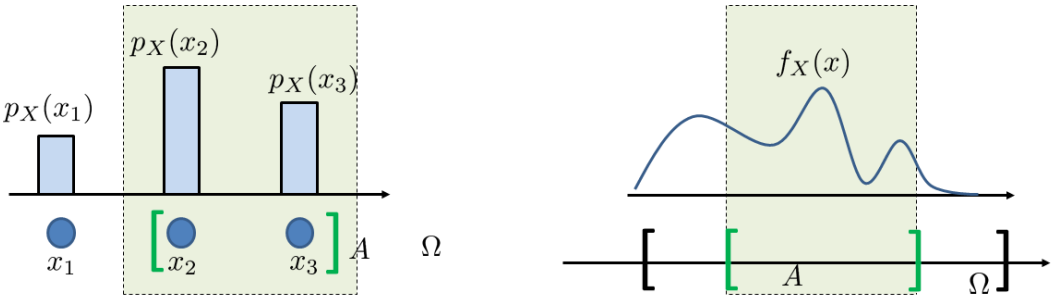


Figure 4.2: [Left] A probability mass function (PMF) tells us the relative frequency of a state when computing the probability. In this example, the "size" of A is $p_X(x_2) + p_X(x_3)$. [Right] A probability density function (PDF) is the infinitesimal version of the PMF. Thus, the "size" of A is the integration over the PDF.

What is a PDF?

- A PDF is the continuous version of a PMF.
- We integrate a PDF to compute the probability.
- We integrate instead of sum because continuous events are not countable.

To summarize, we have learned that when measuring the size of a continuous event, the discrete technique (counting the number of elements) does not work. Generalizing to continuous space requires us to integrate the event. However, since different elements in an event have different relative emphases, we use the probability density function $f_X(x)$ to tell us the relative frequency for a state x to happen. This PDF serves the role of the PMF.

4.1.2 More in-depth discussion about PDFs

A continuous random variable X is defined by its probability density function f_X. This function has to satisfy several criteria, summarized as follows.

Definition 4.1. *A probability density function f_X of a random variable X is a mapping $f_X : \Omega \to \mathbb{R}$, with the properties*

- **Non-negativity**: $f_X(x) \geq 0$ *for all* $x \in \Omega$
- **Unity**: $\int_\Omega f_X(x)\, dx = 1$
- **Measure of a set**: $\mathbb{P}[\{x \in A\}] = \int_A f_X(x)\, dx$

If all elements of the sample space are equiprobable, then the PDF is $f(x) = 1/|\Omega|$. You can easily check that it satisfies all three criteria.

Let us take a closer look at the three criteria:

- Non-negativity: The non-negativity criterion $f_X(x) \geq 0$ is reminiscent of Probability Axiom I. It says that no matter what x we are looking at, the probability density function f_X evaluated at x should never give a negative value. Axiom I ensures that we will not get a negative probability.

- Unity: The unity criterion $\int_\Omega f(x)\ dx = 1$ is reminiscent of Probability Axiom II, which says that measuring over the entire sample space will give 1.
- Measure of a set: The third criterion gives us a way to measure the size of an event A. It says that since each $x \in \Omega$ has a different emphasis when calculating the size of A, we need to scale the elements properly. This scaling is done by the PDF $f_X(x)$, which can be regarded as a histogram with a continuous x-axis. The third criterion is a consequence of Probability Axiom III, because if there are two events A and B that are disjoint, then $\mathbb{P}[\{x \in A\} \cup \{x \in B\}] = \int_A f_X(x)\ dx + \int_B f_X(x)\ dx$ because $f_X(x) \geq 0$ for all x.

If the random variable X takes real numbers in 1D, then a more "user-friendly" definition of the PDF can be given.

Definition 4.2. *Let X be a continuous random variable. The* **probability density function (PDF)** *of X is a function $f_X : \Omega \to \mathbb{R}$ that, when integrated over an interval $[a, b]$, yields the probability of obtaining $a \leq X \leq b$:*

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f_X(x)\ dx. \tag{4.2}$$

This definition is just a rewriting of the previous definition by explicitly writing out the definition of A as an interval $[a, b]$. Here are a few examples.

Example 4.1. Let $f_X(x) = 3x^2$ with $\Omega = [0, 1]$. Let $A = [0, 0.5]$. Then the probability $\mathbb{P}[\{X \in A\}]$ is

$$\mathbb{P}[0 \leq X \leq 0.5] = \int_0^{0.5} 3x^2\ dx = \frac{1}{8}.$$

Example 4.2. Let $f_X(x) = 1/|\Omega|$ with $\Omega = [0, 5]$. Let $A = [3, 5]$. Then the probability $\mathbb{P}[\{X \in A\}]$ is

$$\mathbb{P}[3 \leq X \leq 5] = \int_3^5 \frac{1}{|\Omega|}\ dx = \int_3^5 \frac{1}{5}\ dx = \frac{2}{5}.$$

Example 4.3. Let $f_X(x) = 2x$ with $\Omega = [0, 1]$. Let $A = \{0.5\}$. Then the probability $\mathbb{P}[\{X \in A\}]$ is

$$\mathbb{P}[X = 0.5] = \mathbb{P}[0.5 \leq X \leq 0.5] = \int_{0.5}^{0.5} 2x\ dx = 0.$$

This example shows that evaluating the probability at an isolated point for a continuous random variable will yield 0.

Practice Exercise 4.1. Let X be the phase angle of a voltage signal. Without any prior knowledge about X we may assume that X has an equal probability of any value between 0 to 2π. Find the PDF of X and compute $\mathbb{P}[0 \le X \le \pi/2]$.

Solution. Since X has an equal probability for any value between 0 to 2π, the PDF of X is

$$f_X(x) = \frac{1}{2\pi}, \qquad \text{for } 0 \le x \le 2\pi.$$

Therefore, the probability $\mathbb{P}[0 \le X \le \pi/2]$ can be computed as

$$\mathbb{P}\left[0 \le X \le \frac{\pi}{2}\right] = \int_0^{\pi/2} \frac{1}{2\pi}\, dx = \frac{1}{4}.$$

Looking at Equation (4.2), you may wonder: If the PDF f_X is analogous to PMF p_X, why didn't we require $0 \le f_X(x) \le 1$ instead of requiring only $f_X(x) \ge 0$? This is an excellent question, and it points exactly to the difference between a PMF and a PDF. Notice that f_X is a mapping from the sample space Ω to the real line $\mathbb{R}$. It does not map Ω to $[0, 1]$. On the other hand, since $p_X(x)$ is the actual probability, it maps Ω to $[0, 1]$. Thus, $f_X(x)$ can take very large values but will not explode, because we have the unity constraint $\int_\Omega f_X(x)\, dx = 1$. Even if $f_X(x)$ takes a large value, it will be compensated by the small dx. If you recall, there is nothing like dx in the definition of a PMF. Whenever there is a probability mass, we need to sum or, putting it another way, the dx in the discrete case is always 1. Therefore, while the probability mass PMF must not exceed 1, a probability density PDF can exceed 1.

If $f_X(x) \ge 1$, then what is the meaning of $f_X(x)$? Isn't it representing the probability of having an element $X = x$? If it were a discrete random variable, then yes; $p_X(x)$ is the probability of having $X = x$ (so the probability mass cannot go beyond 1). However, for a continuous random variable, $f_X(x)$ is *not* the probability of having $X = x$. The probability of having $X = x$ (i.e., exactly at x) is 0 because an isolated point has zero measure in the continuous space. Thus, even though $f_X(x)$ takes a value larger than 1, the probability of X being x is zero.

At this point you can see why we call PDF a *density*, or density function, because each value $f_X(x)$ is the probability *per unit length*. If we want to calculate the probability of $x \le X \le x + \delta$, for example, then according to our definition, we have

$$\mathbb{P}[x \le X \le x + \delta] = \int_x^{x+\delta} f_X(x)\, dx \approx f_X(x) \cdot \delta.$$

Therefore, the probability of $\mathbb{P}[x \le X \le x + \delta]$ can be regarded as the "per unit length" density $f_X(x)$ multiplied with the "length" δ. As $\delta \to 0$, we can see that $\mathbb{P}[X = x] = 0$. See **Figure 4.3** for an illustration.

Why are PDFs called a density function?

- Because $f_X(x)$ is the probability **per unit length**.
- You need to integrate $f_X(x)$ to obtain a probability.

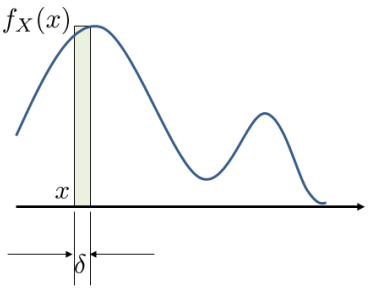


Figure 4.3: The probability $\mathbb{P}[x \le X \le x+\delta]$ can be approximated by the density $f_X(x)$ multiplied by the length δ.

Example 4.4. Consider a random variable X with PDF $f_X(x) = \frac{1}{2\sqrt{x}}$ for any $0 < x \le 1$, and is 0 otherwise. We can show that $f_X(x) \to \infty$ as $x \to 0$. However, $f_X(x)$ remains a valid PDF because

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^1 \frac{1}{2\sqrt{x}}\,dx = \sqrt{x}\Big|_0^1 = 1.$$

Remark. Since isolated points have zero measure in the continuous space, the probability of an open interval (a,b) is the same as the probability of a closed interval:

$$\mathbb{P}[[a,b]] = \mathbb{P}[(a,b)] = \mathbb{P}[(a,b]] = \mathbb{P}[[a,b)].$$

The exception is that when the PDF of $f_X(x)$ has a delta function at a or b. In this case, the probability measure at a or b will be non-zero. We will discuss this when we talk about the CDFs.

Practice Exercise 4.2. Let $f_X(x) = c(1-x^2)$ for $-1 \le x \le 1$, and 0 otherwise. Find the constant c.

Solution. Since $\int_\Omega f_X(x)\,dx = 1$, it follows that

$$\int_\Omega f_X(x)\,dx = \int_{-1}^1 c(1-x^2)\,dx = \frac{4c}{3} \;\Rightarrow\; c = 3/4.$$

Practice Exercise 4.3. Let $f_X(x) = x^2$ for $|x| \le a$, and 0 otherwise. Find a.

Solution. Note that

$$\int_\Omega f_X(x)\,dx = \int_{-a}^a x^2\,dx = \frac{x^3}{3}\Big|_{-a}^a = \frac{2a^3}{3}.$$

Setting $\frac{2a^3}{3} = 1$ yields $a = \sqrt[3]{\frac{3}{2}}$.

4.1.3 Connecting with the PMF

The probability density function is more general than the probability mass function. To see this, consider a discrete random variable X with a PMF $p_X(x)$. Because p_X is defined on a countable set Ω, we can write it as a train of **delta functions** and define a corresponding PDF:

$$f_X(x) = \sum_{x_k \in \Omega} p_X(x_k)\, \delta(x - x_k).$$

Example 4.5. If X is a Bernoulli random variable with PMF $p_X(1) = p$ and $p_X(0) = 1 - p$, then the corresponding PDF can be written as

$$f_X(x) = p\, \delta(x - 1) + (1 - p)\, \delta(x - 0).$$

Example 4.6. If X is a binomial random variable with PMF $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$, then the corresponding PDF can be written as

$$\begin{aligned} f_X(x) &= \sum_{k=0}^{n} p_X(k)\, \delta(x - k) \\ &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}\, \delta(x - k). \end{aligned}$$

Strictly speaking, delta functions are not really functions. They are defined through integrations. They satisfy the properties that $\delta(x - x_k) = \infty$ if $x = x_k$, $\delta(x - x_k) = 0$ if $x \neq x_k$, and

$$\int_{x_k - \epsilon}^{x_k + \epsilon} \delta(x - x_k)\, dx = 1,$$

for any $\epsilon > 0$. Suppose we ignore the fact that delta functions are not functions and merely treat them as ordinary functions with some interesting properties. In this case, we can imagine that for every probability mass $p_X(x_k)$, there exists an interval $[a, b]$ such that there is one and only one state x_k that lies in $[a, b]$, as shown in **Figure 4.4**.

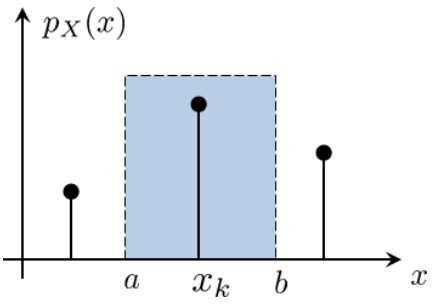


Figure 4.4: We can view a PMF as a train of impulses. When computing the probability $X = x_k$, we integrate the PMF over the interval $[a, b]$.

If we want to calculate the probability of obtaining $X = x_k$, we can show that

$$\begin{aligned}\mathbb{P}[X = x_k] &\overset{(a)}{=} \mathbb{P}[a \le X \le b] \\ &= \int_a^b f_X(x)\, dx \\ &\overset{(b)}{=} \int_a^b p_X(x_k)\, \delta(x - x_k)\, dx \\ &\overset{(c)}{=} p_X(x_k) \underbrace{\int_a^b \delta(x - x_k)\, dx}_{=1} = p_X(x_k).\end{aligned}$$

Here, step (a) holds because within $[a, b]$, there is no other event besides $X = x_k$. Step (b) is just the definition of our $f_X(x)$ (inside the interval $[a, b]$). Step (c) shows that the delta function integrates to 1, thus leaving the probability mass $p_X(x_k)$ as the final result. Let us look at an example and then comment on this intuition.

Example 4.7. Let X be a discrete random variable with PMF

$$p_X(k) = \frac{1}{2^k}, \qquad k = 1, 2, \ldots$$

The continuous representation of the PMF can be written as

$$f_X(x) = \sum_{k=1}^{\infty} p_X(k)\, \delta(x - k) = \sum_{k=1}^{\infty} \left(\frac{1}{2^k}\right) \delta(x - k).$$

Suppose we want to compute the probability $\mathbb{P}[1 \le X \le 2]$. This can be computed as

$$\begin{aligned}\mathbb{P}[1 \le X \le 2] &= \int_1^2 f_X(x)\, dx = \int_1^2 \sum_{k=1}^{\infty} \left(\frac{1}{2^k}\right) \delta(x - k)\, dx \\ &= \int_1^2 \left\{ \frac{1}{2}\delta(x-1) + \frac{1}{4}\delta(x-2) + \cdots \right\} dx \\ &= \frac{1}{2} \underbrace{\int_1^2 \delta(x-1)\, dx}_{=1} + \frac{1}{4} \underbrace{\int_1^2 \delta(x-2)\, dx}_{=1} \\ &\quad + \frac{1}{8} \underbrace{\int_1^2 \delta(x-3)\, dx}_{=0} + \underbrace{\cdots}_{=0} \\ &= \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.\end{aligned}$$

However, if we want to compute the probability $\mathbb{P}[1 < X \le 2]$, then the integration

limit will not include the number 1 and so the delta function will remain 0. Thus,

$$\begin{aligned}\mathbb{P}[1 < X \le 2] &= \int_{1^+}^{2} f_X(x)\,dx \\ &= \frac{1}{2}\underbrace{\int_{1^+}^{2} \delta(x-1)\,dx}_{=0} + \frac{1}{4}\underbrace{\int_{1^+}^{2} \delta(x-2)\,dx}_{=1} = \frac{1}{4}.\end{aligned}$$

Closing remark. To summarize, we see that a PMF can be "regarded" as a PDF. We are careful to put a quotation around "regarded" because PMF and PDF are defined for different events. A PMF uses a discrete measure (i.e., a counter) for countable events, whereas a PDF uses a continuous measure (i.e., integration) for continuous events. The way we link the two is by using the delta functions. Using the delta functions is valid, but the argument we provide here is intuitive rather than rigorous. It is not rigorous because the integration we use is still the Riemann-Stieltjes integration, which does not handle delta functions. Therefore, while you can treat a discrete PDF as a train of delta functions, it is important to remember the limitations of the integrations we use.

4.2 Expectation, Moment, and Variance

4.2.1 Definition and properties

As with discrete random variables, we can define **expectation** for continuous random variables. The definition is analogous: Just replace the summation with integration.

Definition 4.3. *The* **expectation** *of a continuous random variable X is*

$$\mathbb{E}[X] = \int_{\Omega} x\, f_X(x)\,dx. \tag{4.3}$$

Example 4.8. (**Uniform random variable**) Let X be a continuous random variable with PDF $f_X(x) = \frac{1}{b-a}$ for $a \le x \le b$, and 0 otherwise. The expectation is

$$\begin{aligned}\mathbb{E}[X] = \int_{\Omega} x f_X(x)\,dx = \int_a^b x \cdot \frac{1}{b-a}\,dx &= \frac{1}{b-a}\underbrace{\int_a^b x\,dx}_{=\frac{x^2}{2}\Big|_a^b} \\ &= \frac{1}{b-a}\cdot\frac{b^2-a^2}{2} = \frac{a+b}{2}.\end{aligned}$$

Example 4.9. (**Exponential random variable**) Let X be a continuous random variable with PDF $f_X(x) = \lambda e^{-\lambda x}$, for $x \geq 0$. The expectation is

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x\ \lambda e^{-\lambda x}\ dx \\ &= -\int_0^\infty x\ de^{-\lambda x} \\ &= \underbrace{-xe^{-\lambda x}\Big|_0^\infty}_{=0} + \int_0^\infty e^{-\lambda x}\ dx \\ &= -\frac{1}{\lambda}\ \underbrace{e^{-\lambda x}\Big|_0^\infty}_{=-1} = \frac{1}{\lambda},\end{aligned}$$

where the colored step is due to integration by parts.

If a function g is applied to the random variable X, the expectation can be found using the following theorem.

Theorem 4.1. *Let $g : \Omega \to \mathbb{R}$ be a function and X be a continuous random variable. Then*

$$\mathbb{E}[g(X)] = \int_\Omega g(x)\, f_X(x)\ dx. \tag{4.4}$$

Example 4.10. (**Uniform random variable**) Let X be a continuous random variable with $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and 0 otherwise. If $g(\cdot) = (\cdot)^2$, then

$$\begin{aligned}\mathbb{E}[g(X)] = \mathbb{E}[X^2] &= \int_\Omega x^2 f_X(x)\ dx \\ &= \frac{1}{b-a} \cdot \underbrace{\int_a^b x^2\ dx}_{=\frac{b^3-a^3}{3}} = \frac{a^2+ab+b^2}{3}.\end{aligned}$$

Practice Exercise 4.4. Let Θ be a continuous random variable with PDF $f_\Theta(\theta) = \frac{1}{2\pi}$ for $0 \leq \theta \leq 2\pi$ and is 0 otherwise. Let $Y = \cos(\omega t + \Theta)$. Find $\mathbb{E}[Y]$.

Solution. Referring to Equation (4.4), the function g is

$$g(\theta) = \cos(\omega t + \theta).$$

Therefore, the expectation $\mathbb{E}[Y]$ is

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^{2\pi} \cos(\omega t + \theta)\ f_\Theta(\theta)\ d\theta \\ &= \frac{1}{2\pi}\int_0^{2\pi} \cos(\omega t + \theta)\ d\theta = 0,\end{aligned}$$

where the last equality holds because the integral of a sinusoid over one period is 0.

Practice Exercise 4.5. Let $A \subseteq \Omega$. Let $\mathbb{I}_A(X)$ be an indicator function such that

$$\mathbb{I}_A(X) = \begin{cases} 1, & \text{if } X \in A, \\ 0, & \text{if } X \notin A. \end{cases}$$

Find $\mathbb{E}[\mathbb{I}_A(X)]$.

Solution. The expectation is

$$\mathbb{E}[\mathbb{I}_A(X)] = \int_\Omega \mathbb{I}_A(x) f_X(x)\ dx = \int_{x\in A} f_X(x)\ dx = \mathbb{P}[X \in A].$$

So the probability of $\{X \in A\}$ can be equivalently represented in terms of expectation.

Practice Exercise 4.6. Is it true that $\mathbb{E}[1/X] = 1/\mathbb{E}[X]$?

Solution. No. This is because

$$\mathbb{E}\left[\frac{1}{X}\right] = \int_\Omega \frac{1}{x} f_X(x)\ dx \neq \frac{1}{\int_\Omega x f_X(x)\ dx} = \frac{1}{\mathbb{E}[X]}.$$

All the properties of expectation we learned in the discrete case can be translated to the continuous case. Specifically, we have that

- $\mathbb{E}[aX] = a\mathbb{E}[X]$: A scalar multiple of a random variable will scale the expectation.
- $\mathbb{E}[X+a] = \mathbb{E}[X]+a$: Constant addition of a random variable will offset the expectation.
- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$: Affine transformation of a random variable will translate to the expectation.

Practice Exercise 4.7. Prove the above three statements.

Solution. The third statement is just the sum of the first two statements, so we just

need to show the first two:

$$\mathbb{E}[aX] = \int_\Omega axf_X(x)\,dx = a\int_\Omega xf_X(x)\,dx = a\mathbb{E}[X],$$
$$\mathbb{E}[X+a] = \int_\Omega (x+a)f_X(x)\,dx = \int_\Omega xf_X(x)\,dx + a = \mathbb{E}[X] + a.$$

4.2.2 Existence of expectation

As we discussed in the discrete case, not all random variables have an expectation.

Definition 4.4. *A random variable X has an expectation if it is* **absolutely integrable**, *i.e.,*

$$\mathbb{E}[|X|] = \int_\Omega |x| f_X(x)\,dx < \infty. \tag{4.5}$$

Being absolutely integrable implies that the expectation is that $\mathbb{E}[|X|]$ is the upper bound of $\mathbb{E}[X]$.

Theorem 4.2. *For any random variable X,*

$$|\mathbb{E}[X]| \le \mathbb{E}[|X|]. \tag{4.6}$$

Proof. Note that $f_X(x) \ge 0$. Therefore,

$$-|x|\, f_X(x) \le x\, f_X(x) \le |x|, f_X(x), \qquad \forall x.$$

Thus, integrating all three terms yields

$$-\int_\Omega |x| f_X(x)\,dx \le \int_\Omega x\, f_X(x)\,dx \le \int_\Omega |x| f_X(x)\,dx,$$

which is equivalent to $-\mathbb{E}[|X|] \le \mathbb{E}[X] \le \mathbb{E}[|X|]$.

□

Example 4.11. Here is a random variable whose expectation is undefined. Let X be a random variable with PDF

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \qquad x \in \mathbb{R}.$$

This random variable is called the **Cauchy random variable**. We can show that

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot \frac{1}{\pi(1+x^2)}\,dx = \frac{1}{\pi}\int_0^\infty \frac{x}{(1+x^2)}\,dx + \frac{1}{\pi}\int_{-\infty}^0 \frac{x}{(1+x^2)}\,dx.$$

The first integral gives

$$\int_0^\infty \frac{x}{(1+x^2)}\,dx = \frac{1}{2}\log(1+x^2)\Big|_0^\infty = \infty,$$

and the second integral gives $-\infty$. Since neither integral is finite, the expectation is undefined. We can also check the absolutely integrability criterion:

$$\begin{aligned}\mathbb{E}[|X|] &= \int_{-\infty}^{\infty} |x| \cdot \frac{1}{\pi(1+x^2)}\,dx \\ &\overset{(a)}{=} 2\int_0^\infty \frac{x}{\pi(1+x^2)}\,dx \geq 2\int_1^\infty \frac{x}{\pi(1+x^2)}\,dx \\ &\overset{(b)}{\geq} 2\int_1^\infty \frac{x}{\pi(x^2+x^2)}\,dx = \frac{1}{\pi}\log(x)\Big|_1^\infty = \infty,\end{aligned}$$

where in (a) we use the fact that the function being integrated is even, and in (b) we lower-bound $\frac{1}{1+x^2} \geq \frac{1}{x^2+x^2}$ if $x > 1$.

4.2.3 Moment and variance

The moment and variance of a continuous random variable can be defined analogously to the moment and variance of a discrete random variable, replacing the summations with integrations.

Definition 4.5. *The* ***k*th moment** *of a continuous random variable X is*

$$\mathbb{E}[X^k] = \int_\Omega x^k\, f_X(x)\,dx. \tag{4.7}$$

Definition 4.6. *The* **variance** *of a continuous random variable X is*

$$\mathrm{Var}[X] = \mathbb{E}[(X-\mu)^2] = \int_\Omega (x-\mu)^2 f_X(x)\,dx, \tag{4.8}$$

where $\mu \overset{\text{def}}{=} \mathbb{E}[X]$.

It is not difficult to show that the variance can also be expressed as

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mu^2,$$

because

$$\mathrm{Var}[X] = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - 2\mathbb{E}[X]\mu + \mu^2 = \mathbb{E}[X^2] - \mu^2.$$

Practice Exercise 4.8. (**Uniform random variable**) Let X be a continuous random variable with PDF $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and 0 otherwise. Find $\text{Var}[X]$.

Solution. We have shown that $\mathbb{E}[X] = \frac{a+b}{2}$ and $\mathbb{E}[X^2] = \frac{a^2+ab+b^2}{3}$. Therefore, the variance is

$$\begin{aligned}\text{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{(b-a)^2}{12}.\end{aligned}$$

Practice Exercise 4.9. (**Exponential random variable**) Let X be a continuous random variable with PDF $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and 0 otherwise. Find $\text{Var}[X]$.

Solution. We have shown that $\mathbb{E}[X] = \frac{1}{\lambda}$. The second moment is

$$\begin{aligned}\mathbb{E}[X^2] &= \int_0^\infty x^2 \, \lambda e^{-\lambda x} \, dx \\ &= \left[-x^2 e^{-\lambda x}\right]_0^\infty + \int_0^\infty 2xe^{-\lambda x} \, dx \\ &= \frac{2}{\lambda} \int_0^\infty x\lambda e^{-\lambda x} \, dx \\ &= \frac{2}{\lambda} \cdot \frac{1}{\lambda} = \frac{2}{\lambda^2}.\end{aligned}$$

Therefore,

$$\begin{aligned}\text{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.\end{aligned}$$

4.3 Cumulative Distribution Function

When we discussed discrete random variables, we introduced the concept of cumulative distribution functions (CDFs). One of the motivations was that if we view a PMF as a train of delta functions, they are technically not well-defined functions. However, it turns out that the CDF is always a well-defined function. In this section, we will complete the story by first discussing the CDF for continuous random variables. Then, we will come back and show you how the CDF can be derived for discrete random variables.

4.3.1 CDF for continuous random variables

Definition 4.7. *Let X be a continuous random variable with a sample space $\Omega = \mathbb{R}$. The* **cumulative distribution function (CDF)** *of X is*

$$F_X(x) \stackrel{\text{def}}{=} \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f_X(x')\, dx'. \tag{4.9}$$

The interpretation of the CDF can be seen from **Figure 4.5**. Given a PDF f_X, the CDF F_X evaluated at x is the integration of f_X from $-\infty$ up to a point x. The integration of f_X from $-\infty$ to x is nothing but the area under the curve of f_X. Since f_X is non-negative, the larger value x we use to evaluate in $F_X(x)$, the more area under the curve we are looking at. In the extreme when $x = -\infty$, we can see that $F_X(-\infty) = 0$, and when $x = +\infty$ we have that $F_X(+\infty) = \int_{-\infty}^{\infty} f_X(x)\, dx = 1$.

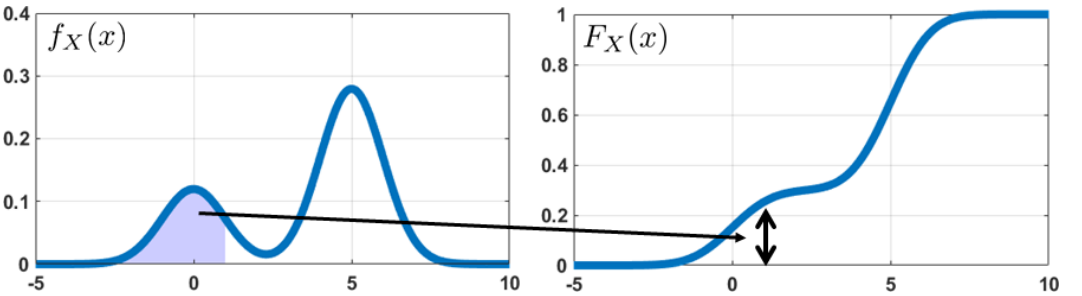


Figure 4.5: A CDF is the integral of the PDF. Thus, the height of a stem in the CDF corresponds to the area under the curve of the PDF.

Practice Exercise 4.10. (**Uniform random variable**) Let X be a continuous random variable with PDF $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and is 0 otherwise. Find the CDF of X.

Solution. The CDF of X is given by

$$F_X(x) = \begin{cases} 0, & x \leq a, \\ \int_{-\infty}^{x} f_X(x')\, dx' = \int_a^x \frac{1}{b-a}\, dx' = \frac{x-a}{b-a}, & a < x \leq b, \\ 1, & x > b. \end{cases}$$

As you can see from this practice exercise, we explicitly break the CDF into three segments. The first segment gives $F_X(x) = 0$ because for any $x \leq a$, there is nothing to integrate, since $f_X(x) = 0$ for any $x \leq a$. Similarly, for the last segment, $F_X(x) = 1$ for all $x > b$ because once x goes beyond b, the integration will cover all the non-zeros of f_X. **Figure 4.6** illustrates the PDF and CDF for this example.

In MATLAB, we can generate the PDF and CDF using the commands `pdf` and `cdf` respectively. For the particular example shown in **Figure 4.6**, the following code can be used. A similar set of commands can be implemented in Python.

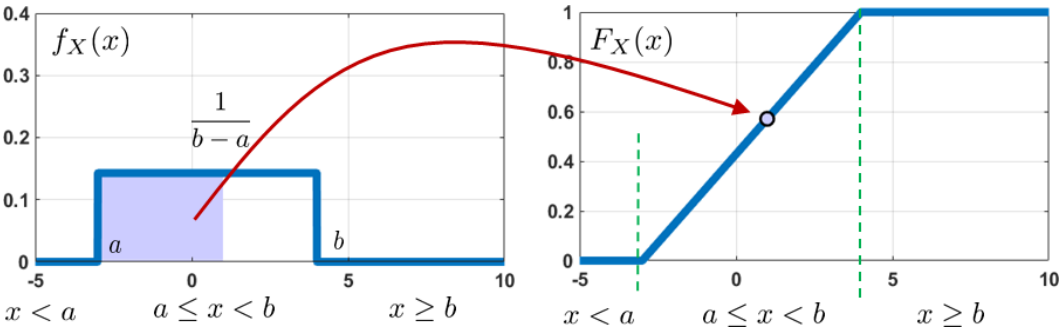


Figure 4.6: Example: $f_X(x) = 1/(b-a)$ for $a \le x \le b$. The CDF has three segments.

```
% MATLAB code to generate the PDF and CDF
unif = makedist('Uniform','lower',-3,'upper',4);
x  = linspace(-5, 10, 1500)';
f  = pdf(unif, x);
F  = cdf(unif, x);
figure(1); plot(x, f, 'LineWidth', 6);
figure(2); plot(x, F, 'LineWidth', 6);
```

```
# Python code to generate the PDF and CDF
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
x = np.linspace(-5,10,1500)
f = stats.uniform.pdf(x,-3,4)
F = stats.uniform.cdf(x,-3,4)
plt.plot(x,f); plt.show()
plt.plot(x,F); plt.show()
```

Practice Exercise 4.11. (**Exponential random variable**) Let X be a continuous random variable with PDF $f_X(x) = \lambda e^{-\lambda x}$ for $x \ge 0$, and 0 otherwise. Find the CDF of X.

Solution. Clearly, for $x < 0$, we have $F_X(x) = 0$. For $x \ge 0$, we can show that

$$F_X(x) = \int_0^x f_X(x')\, dx' = \int_0^x \lambda e^{-\lambda x'}\, dx' = 1 - e^{-\lambda x}.$$

Therefore, the complete CDF is (see **Figure 4.7** for illustration):

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \ge 0. \end{cases}$$

The MATLAB code and Python code to generate this figure are shown below.

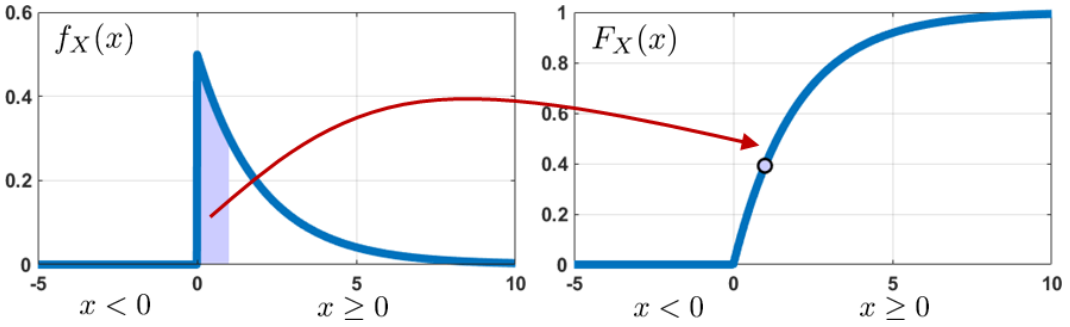


Figure 4.7: Example: $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. The CDF has two segments.

```
% MATLAB code to generate the PDF and CDF
pd = makedist('exp',2);
x  = linspace(-5, 10, 1500)';
f  = pdf(pd, x);
F  = cdf(pd, x);
figure(1); plot(x, f, 'LineWidth', 6);
figure(2); plot(x, F, 'LineWidth', 6);
```

```
# Python code to generate the PDF and CDF
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
x = np.linspace(-5,10,1500)
f = stats.expon.pdf(x,2)
F = stats.expon.cdf(x,2)
plt.plot(x,f); plt.show()
plt.plot(x,F); plt.show()
```

4.3.2 Properties of CDF

Let us now describe the properties of a CDF. If we compare these with those for the discrete cases, we see that the continuous cases simply replace the summations by integrations. Therefore, we should expect to inherit most of the properties from the discrete cases.

Proposition 4.1. *Let X be a random variable (either continuous or discrete), then the CDF of X has the following properties:*

(i) The CDF is **nondecreasing**.

(ii) The **maximum** *of the CDF is when $x = \infty$: $F_X(+\infty) = 1$.*

(iii) The **minimum** *of the CDF is when $x = -\infty$: $F_X(-\infty) = 0$.*

Proof. For (i), we notice that $F_X(x) = \int_{-\infty}^{x} f_X(x')\,dx'$. Therefore, if $s \le t$ then

$$F_X(s) = \int_{-\infty}^{s} f_X(x')\,dx' \le \int_{-\infty}^{t} f_X(x')\,dx' = F_X(t).$$

Thus it shows that F_X is nondecreasing. (It does not need to be increasing because a CDF can have a steady state.) For (ii) and (iii), we can show that

$$F_X(+\infty) = \int_{-\infty}^{+\infty} f_X(x')\,dx' = 1, \quad \text{and} \quad F_X(-\infty) = \int_{-\infty}^{-\infty} f_X(x')\,dx' = 0. \qquad \square$$

Example 4.12. We can show that the CDF we derived for the uniform random variable satisfies these three properties. To see this, we note that

$$F_X(x) = \frac{x-a}{b-a}, \quad a \le x \le b.$$

The derivative of this function $F_X'(x) = \frac{1}{b-a} > 0$ for $a \le x \le b$. Also, note that $F_X(x) = 0$ for $x < a$ and $x > b$, so F_X is nondecreasing. The other two properties follow because if $x = b$, then $F_X(b) = 1$, and if $x = a$ then $F_X(a) = 0$. Together with the nondecreasing property, we show (ii) and (iii).

Proposition 4.2. *Let X be a continuous random variable. If the CDF F_X is continuous at any $a \le x \le b$, then*

$$\mathbb{P}[a \le X \le b] = F_X(b) - F_X(a). \tag{4.10}$$

Proof. The proof follows from the definition of the CDF, which states that

$$\begin{aligned} F_X(b) - F_X(a) &= \int_{-\infty}^{b} f_X(x')\,dx' - \int_{-\infty}^{a} f_X(x')\,dx' \\ &= \int_{a}^{b} f_X(x')\,dx' = \mathbb{P}[a \le X \le b]. \qquad \square \end{aligned}$$

This result provides a very handy tool for calculating the probability of an event $a \le X \le b$ using the CDF. It says that $\mathbb{P}[a \le X \le b]$ is the difference between $F_X(b)$ and $F_X(a)$. So, if we are given F_X, calculating the probability of $a \le X \le b$ just involves evaluating the CDF at a and b. The result also shows that for a continuous random variable X, $\mathbb{P}[X = x_0] = F_X(x_0) - F_X(x_0) = 0$. This is consistent with our arguments from the measure's point of view.

Example 4.13. (**Exponential random variable**) We showed that the exponential random variable X with a PDF $f_X(x) = \lambda e^{-\lambda x}$ for $x \ge 0$ (and $f_X(x) = 0$ for $x < 0$) has a CDF given by $F_X(x) = 1 - e^{-\lambda x}$ for $x \ge 0$. Suppose we want to calculate the

probability $\mathbb{P}[1 \le X \le 3]$. Then the PDF approach gives us

$$\mathbb{P}[1 \le X \le 3] = \int_1^3 f_X(x)\,dx = \int_1^3 \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}\Big|_1^3 = e^{-3\lambda} - e^{-\lambda}.$$

If we take the CDF approach, we can show that

$$\begin{aligned}\mathbb{P}[1 \le X \le 3] &= F_X(3) - F_X(1) \\ &= (1 - e^{-\lambda}) - (1 - e^{-3\lambda}) = e^{-3\lambda} - e^{-\lambda},\end{aligned}$$

which yields the same as the PDF approach.

Example 4.14. Let X be a random variable with PDF $f_X(x) = 2x$ for $0 \le x \le 1$, and is 0 otherwise. We can show that the CDF is

$$F_X(x) = \int_0^x f_X(t)\,dt = \int_0^x 2t\;\,dt = t^2\Big|_0^x = x^2, \qquad 0 \le x \le 1.$$

Therefore, to compute the probability $\mathbb{P}[1/3 \le X \le 1/2]$, we have

$$\mathbb{P}\left[\frac{1}{3} \le X \le \frac{1}{2}\right] = F_X\left(\frac{1}{2}\right) - F_X\left(\frac{1}{3}\right) = \left(\frac{1}{2}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{5}{36}.$$

□

A CDF can be used for both continuous and discrete random variables. However, before we can do that, we need a tool to handle the discontinuities. The following definition is a summary of the three types of continuity.

Definition 4.8. *A function $F_X(x)$ is said to be*

- **Left-continuous** *at $x = b$ if $F_X(b) = F_X(b^-) \stackrel{\text{def}}{=} \lim_{h\to 0} F_X(b-h)$;*
- **Right-continuous** *at $x = b$ if $F_X(b) = F_X(b^+) \stackrel{\text{def}}{=} \lim_{h\to 0} F_X(b+h)$;*
- **Continuous** *at $x = b$ if it is both right-continuous and left-continuous at $x = b$. In this case, we have*

$$\lim_{h\to 0} F_X(b-h) = \lim_{h\to 0} F_X(b+h) = F(b).$$

In this definition, the step size $h > 0$ is shrinking to zero. The point $b-h$ stays at the left of b, and $b+h$ stays at the right of b. Thus, if we set the limit $h \to 0$, $b-h$ will approach a point b^- whereas $b+h$ will approach a point b^+. If it happens that $F_X(b^-) = F_X(b)$ then we say that F_X is left-continuous at b. If $F_X(b^+) = F_X(b)$ then we say that F_X is right-continuous at b. These are summarized in **Figure 4.8**.

Whenever F_X has a discontinuous point, it can be left-continuous, right-continuous, or neither. ("Neither" happens if $F_X(b)$ take a value other than $F_X(b^+)$ or $F_X(b^-)$. You can

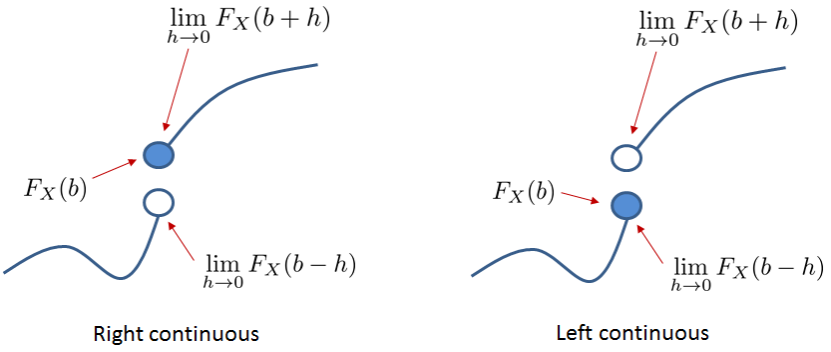


Figure 4.8: The definition of left- and right-continuous at a point b.

always create a nasty function that satisfies this condition.) For continuous functions, it is necessary that $F_X(b^-) = F_X(b^+)$. If this happens, there is no gap between the two points.

Theorem 4.3. *For any random variable X (discrete or continuous), $F_X(x)$ is always* **right-continuous**. *That is,*

$$F_X(b) = F_X(b^+) \stackrel{\text{def}}{=} \lim_{h\to 0} F_X(b+h) \tag{4.11}$$

Right-continuous means that if $F_X(x)$ is piecewise, it must have a **solid left end and an empty right end**. Figure 4.9 shows an example of a valid CDF and an invalid CDF.

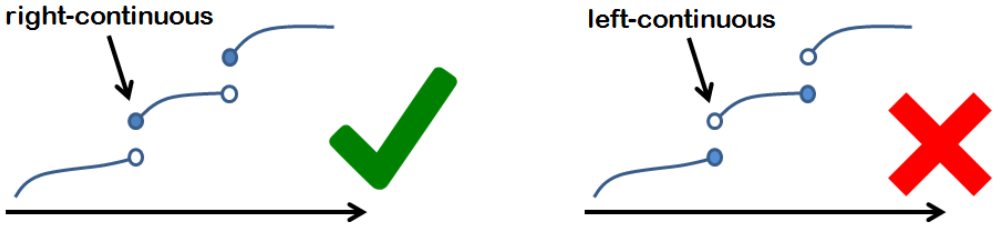


Figure 4.9: A CDF must be right-continuous.

The reason why F_X is always right-continuous is that the inequality $X \leq x$ has a closed right-hand limit. Imagine the following situation: A discrete random variable X has four states: $1, 2, 3, 4$. Then,

$$\lim_{h\to 0} F_X(3+h) = \lim_{h\to 0} \sum_{k=1}^{\text{“}3+h\text{”}} p_X(k) = p_X(1) + p_X(2) + p_X(3) = F_X(3).$$

Similarly, if you have a continuous random variable X with a PDF f_X, then

$$\lim_{h\to 0} F_X(b+h) = \lim_{h\to 0} \int_{-\infty}^{b+h} f_X(t)\, dt = \int_{-\infty}^{b} f_X(t)\, dt = F_X(b).$$

In other words, the "$\leq$" ensures that the rightmost state is included. If we defined CDF using $<$, we would have gotten left-hand continuous, but this would be inconvenient because the $<$ requires us to deal with limits whenever we evaluate $X < x$.

Theorem 4.4. *For any random variable X (discrete or continuous), $\mathbb{P}[X = b]$ is*

$$\mathbb{P}[X = b] = \begin{cases} F_X(b) - F_X(b^-), & \text{if } F_X \text{ is discontinuous at } x = b \\ 0, & \text{otherwise.} \end{cases} \tag{4.12}$$

This proposition states that when $F_X(x)$ is discontinuous at $x = b$, then $\mathbb{P}[X = b]$ is the difference between $F_X(b)$ and the limit from the left. In other words, the height of the gap determines the probability at the discontinuity. If $F_X(x)$ is continuous at $x = b$, then $F_X(b) = \lim_{h \to 0} F_X(b - h)$ and so $\mathbb{P}[X = b] = 0$.

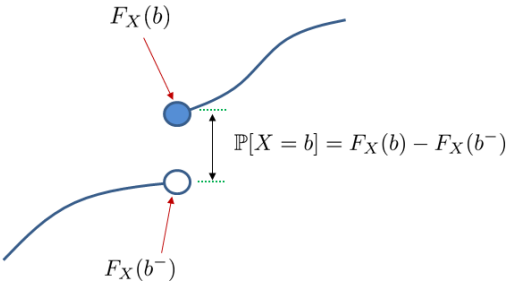


Figure 4.10: Illustration of Equation (4.12). Since the CDF is discontinuous at a point $x = b$, the gap $F_X(b) - F_X(b^-)$ will define the probability $\mathbb{P}[X = b]$.

Example 4.15. Consider a random variable X with a PDF

$$f_X(x) = \begin{cases} x, & 0 \leq x \leq 1, \\ \frac{1}{2}, & x = 3, \\ 0, & \text{otherwise.} \end{cases}$$

The CDF $F_X(x)$ will consist of a few segments. The first segment is $0 \leq x < 1$. We can show that

$$F_X(x) = \int_0^x f_X(t)\, dt = \int_0^x t \;\, dt = \frac{t^2}{2}\bigg|_0^x = \frac{x^2}{2}, \quad 0 \leq x < 1.$$

The second segment is when $1 \leq x < 3$. Since there is no new f_X to integrate, the CDF stays at $F_X(x) = F_X(1) = \frac{1}{2}$ for $1 \leq x < 3$. The third segment is $x > 3$. Because this range has covered the entire sample space, we have $F_X(x) = 1$ for $x > 3$. How about $x = 3$? We can show that

$$F_X(3) = F_X(3^+) = 1.$$

Therefore, to summarize, the CDF is

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \frac{x^2}{2}, & 0 \le x < 1, \\ \frac{1}{2}, & 1 \le x < 3, \\ 1, & x \ge 3. \end{cases}$$

A graphical illustration is shown in **Figure 4.11**.

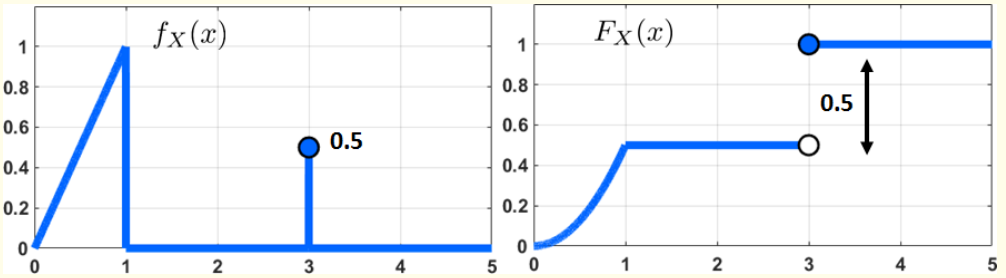


Figure 4.11: An example of converting a PDF to a CDF.

4.3.3 Retrieving PDF from CDF

Thus far, we have only seen how to obtain $F_X(x)$ from $f_X(x)$. In order to go in the reverse direction, we recall the fundamental theorem of calculus. This states that if a function f is continuous, then

$$f(x) = \frac{d}{dx} \int_a^x f(t)\, dt$$

for some constant a. Using this result for CDF and PDF, we have the following:

Theorem 4.5. *The* **probability density function** *(PDF) is the derivative of the cumulative distribution function (CDF):*

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{d}{dx} \int_{-\infty}^x f_X(x')\, dx', \tag{4.13}$$

provided F_X is differentiable at x. If F_X is not differentiable at $x = x_0$, then,

$$f_X(x_0) = \mathbb{P}[X = x_0]\delta(x - x_0).$$

Example 4.16. Consider a CDF

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - \frac{1}{4}e^{-2x}, & x \ge 0. \end{cases}$$

We want to find the PDF $f_X(x)$. To do so, we first show that $F_X(0) = \frac{3}{4}$. This

corresponds to a discontinuity at $x = 0$, as shown in **Figure 4.12**.

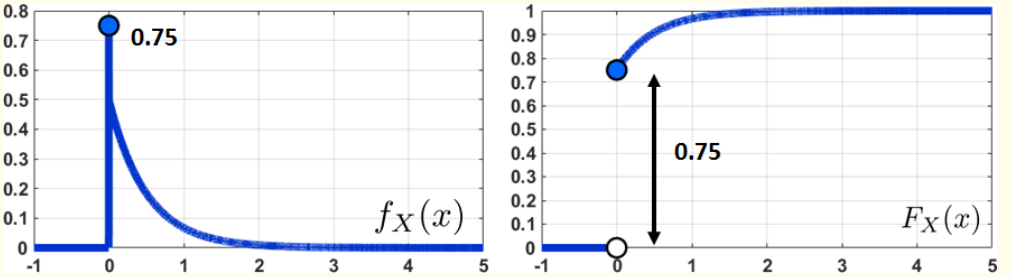


Figure 4.12: An example of converting a PDF to a CDF.

Because of the discontinuity, we need to consider three cases:

$$f_X(x) = \begin{cases} \frac{dF_X(x)}{dx}, & x < 0, \\ \mathbb{P}[X = 0]\ \delta(x - 0), & x = 0, \\ \frac{dF_X(x)}{dx}, & x > 0. \end{cases}$$

When $x < 0$, $F_X(x) = 0$, so $\frac{dF_X(x)}{dx} = 0$. When $x > 0$, $F_X(x) = 1 - \frac{1}{4}e^{-2x}$, so

$$\frac{dF_X(x)}{dx} = \frac{1}{2}e^{-2x}.$$

When $x = 0$, the probability $\mathbb{P}[X = 0]$ is determined by the gap between the solid dot and the empty dot. This yields

$$\mathbb{P}[X = 0] = F_X(0) - \lim_{h \to 0} F_X(0 - h) = \frac{3}{4} - 0 = \frac{3}{4}.$$

Therefore, the overall PDF is

$$f_X(x) = \begin{cases} 0, & x < 0, \\ \frac{3}{4}\delta(x - 0), & x = 0, \\ \frac{1}{2}e^{-2x}, & x > 0. \end{cases}$$

Figure 4.12 illustrates this example.

4.3.4 CDF: Unifying discrete and continuous random variables

The CDF is always a well-defined function. It is integrable everywhere. If the underlying random variable is continuous, the CDF is also continuous. If the underlying random variable is discrete, the CDF is a staircase function. We have seen enough CDFs for continuous random variables. Let us (re)visit a few discrete random variables.

Example 4.17. (Geometric random variable) Consider a geometric random variable with PMF $p_X(k) = (1 - p)^{k-1}p$, for $k = 1, 2, \ldots$.

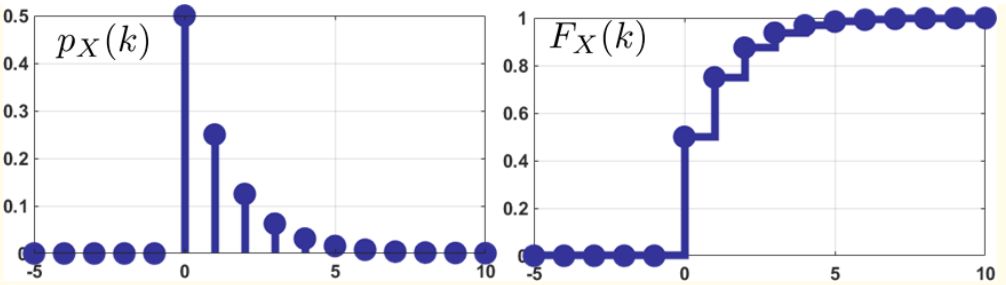


Figure 4.13: PMF and CDF of a geometric random variable.

We can show that the CDF is

$$F_X(k) = \sum_{\ell=1}^{k} p_X(\ell) = \sum_{\ell=1}^{k} (1-p)^{\ell-1} p = p \cdot \frac{1-(1-p)^k}{1-(1-p)} = 1-(1-p)^k.$$

For a sanity check, we can try to retrieve the PMF from the CDF:

$$\begin{aligned} p_X(k) &= F_X(k) - F_X(k-1) \\ &= (1-(1-p)^k) - (1-(1-p)^{k-1}) \\ &= (1-p)^{k-1} p. \end{aligned}$$

A graphical portrayal of this example is shown in **Figure 4.13**.

If we treat the PMFs as delta functions in the above example, then the continuous definition also applies. Since the CDF is a piecewise constant function, the derivative is exactly a delta function. For some problems, it is easier to start with CDF and then compute the PMF or PDF. Here is an example.

Example 4.18. Let X_1, X_2 and X_3 be three independent discrete random variables with sample space $\Omega = \{1, 2, \ldots, 10\}$. Define $X = \max\{X_1, X_2, X_3\}$. We want to find the PMF of X. To tackle this problem, we first observe that the PMF for X_1 is $p_{X_1}(k) = \frac{1}{10}$. Thus, the CDF of X_1 is

$$F_{X_1}(k) = \sum_{\ell=1}^{k} p_{X_1}(\ell) = \frac{k}{10}.$$

Then, we can show that the CDF of X is

$$\begin{aligned} F_X(k) = \mathbb{P}[X \le k] &= \mathbb{P}[\max\{X_1, X_2, X_3\} \le k] \\ &\overset{(a)}{=} \mathbb{P}[X_1 \le k \ \cap X_2 \le k \ \cap X_3 \le k] \\ &\overset{(b)}{=} \mathbb{P}[X_1 \le k]\mathbb{P}[X_2 \le k]\mathbb{P}[X_3 \le k] \\ &= \left(\frac{k}{10}\right)^3, \end{aligned}$$

where in (a) we use the fact that $\max\{X_1, X_2, X_3\} \leq k$ if and only if all three elements are less than k, and in (b) we use independence. Consequently, the PMF of X is

$$p_X(k) = F_X(k) - F_X(k-1) = \left(\frac{k}{10}\right)^3 - \left(\frac{k-1}{10}\right)^3.$$

What is a CDF?

- CDF is $F_X(x) = \mathbb{P}[X \leq x]$. It is the cumulative sum of the PMF/PDF.
- CDF is either a staircase function, a smooth function, or a hybrid. Unlike a PDF, which is not defined for discrete random variables, the CDF is always well defined.
- CDF $\xrightarrow{\frac{d}{dx}}$ PDF.
- CDF $\xleftarrow{\int}$ PDF.
- Gap of jump in CDF = height of delta in PDF.

4.4 Median, Mode, and Mean

There are three statistical quantities that we are frequently interested in: mean, mode, and median. We all know how to compute these from a dataset. For example, to compute the median of a dataset, we sort the data and pick the number that sits in the 50th percentile. However, the median computed in this way is the **empirical median**, i.e., it is a value computed from a particular dataset. If the data is generated from a random variable (with a given PDF), how do we compute the mean, median, and mode?

4.4.1 Median

Imagine you have a sequence of numbers as shown below.

n	1	2	3	4	5	6	7	8	9	$\cdots$	100
x_n	1.5	2.5	3.1	1.1	-0.4	-4.1	0.5	2.2	-3.4	$\cdots$	-1.4

How do we compute the median? We first sort the sequence (either in ascending order or descending order), and then pick the middle one. On computer, we permute the samples

$$\{x_{1'}, x_{2'}, \ldots, x_{N'}\} = \text{sort}\{x_1, x_2, \ldots, x_N\},$$

such that $x_{1'} < x_{2'} < \ldots < x_{N'}$ is ordered. The median is the one positioned at the middle. There are, of course, built-in commands such as `median` in MATLAB and `np.median` in Python to perform the median operation.

Now, how do we compute the median if we are given a random variable X with a PDF $f_X(x)$? The answer is by integrating the PDF.

Definition 4.9. *Let X be a continuous random variable with PDF f_X. The* **median** *of X is a point $c \in \mathbb{R}$ such that*

$$\int_{-\infty}^{c} f_X(x)\,dx = \int_{c}^{\infty} f_X(x)\,dx. \tag{4.14}$$

Why is the median defined in this way? This is because $\int_{-\infty}^{c} f_X(x)\,dx$ is the area under the curve on the left of c, and $\int_{c}^{\infty} f_X(x)\,dx$ is the area under the curve on the right of c. The area under the curve tells us the percentage of numbers that are less than the cutoff. Therefore, if the left area equals the right area, then c must be the median.

How to find the median from the PDF

- Find a point c that separates the PDF into two equal areas

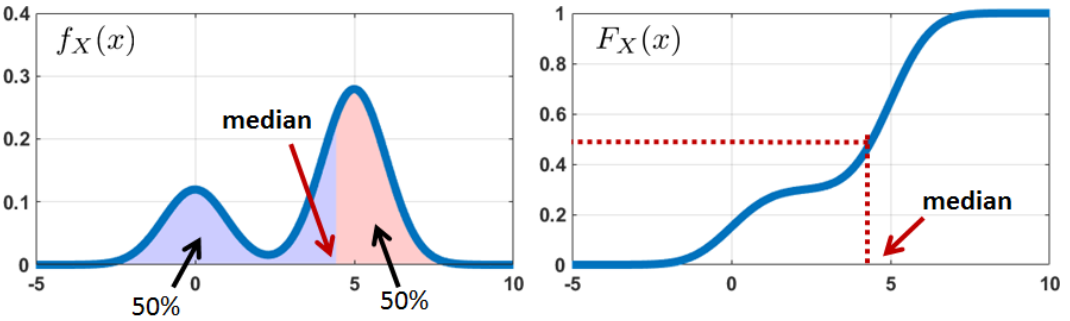


Figure 4.14: [Left] The median is computed as the point such that the two areas under the curve are equal. [Right] The median is computed as the point such that F_X hits 0.5.

The median can also be evaluated from the CDF as follows.

Theorem 4.6. *The* **median** *of a random variable X is the point c such that*

$$F_X(c) = \frac{1}{2}. \tag{4.15}$$

Proof. Since $F_X(x) = \int_{-\infty}^{x} f_X(x')\,dx'$, we have

$$\begin{aligned} F_X(c) &= \int_{-\infty}^{c} f_X(x)\,dx \\ &= \int_{c}^{\infty} f_X(x)\,dx = 1 - F_X(c). \end{aligned}$$

Rearranging the terms shows that $F_X(c) = \frac{1}{2}$. □

How to find median from CDF

- Find a point c such that $F_X(c) = 0.5$.

Example 4.19. (**Uniform random variable**) Let X be a continuous random variable with PDF $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and is 0 otherwise. We know that the CDF of X is $F_X(x) = \frac{x-a}{b-a}$ for $a \leq x \leq b$. Therefore, the median of X is the number $c \in \mathbb{R}$ such that $F_X(c) = \frac{1}{2}$. Substituting into the CDF yields $\frac{c-a}{b-a} = \frac{1}{2}$, which gives $c = \frac{a+b}{2}$.

Example 4.20. (**Exponential random variable**) Let X be a continuous random variable with PDF $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. We know that the CDF of X is $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$. The median of X is the point c such that $F_X(c) = \frac{1}{2}$. This gives $1 - e^{-\lambda c} = \frac{1}{2}$, which is $c = \frac{\log 2}{\lambda}$.

4.4.2 Mode

The mode is the peak of the PDF. We can see this from the definition below.

Definition 4.10. *Let X be a continuous random variable. The mode is the point c such that $f_X(x)$ attains the maximum:*

$$c = \underset{x \in \Omega}{argmax} \; f_X(x) = \underset{x \in \Omega}{argmax} \; \frac{d}{dx} F_X(x). \tag{4.16}$$

The second equality holds because $f_X(x) = F'_X(x) = \frac{d}{dx} \int_{-\infty}^{x} f_X(t)\, dt$. A pictorial illustration of mode is given in **Figure 4.15**. Note that the mode of a random variable is not unique, e.g., a mixture of two identical Gaussians with different means has two modes.

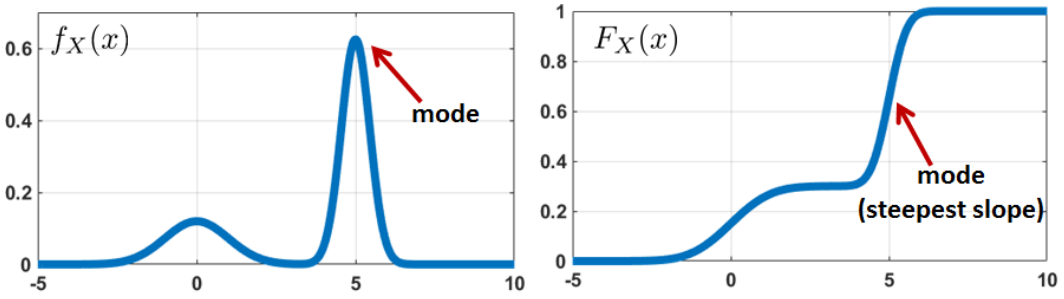


Figure 4.15: [Left] The mode appears at the peak of the PDF. [Right] The mode appears at the steepest slope of the CDF.

How to find mode from PDF

- Find a point c such that $f_X(c)$ is maximized.

How to find mode from CDF

- Continuous: Find a point c such that $F_X(c)$ has the steepest slope.
- Discrete: Find a point c such that $F_X(c)$ has the biggest gap in a jump.

Example 4.21. Let X be a continuous random variable with PDF $f_X(x) = 6x(1-x)$ for $0 \le x \le 1$. The mode of X happens at $\operatorname{argmax}_x f_X(x)$. To find this maximum, we take the derivative of f_X. This gives

$$0 = \frac{d}{dx} f_X(x) = \frac{d}{dx} 6x(1-x) = 6(1-2x).$$

Setting this equal to zero yields $x = \frac{1}{2}$.

To ensure that this point is a maximum, we take the second-order derivative:

$$\frac{d^2}{dx^2} f_X(x) = \frac{d}{dx} 6(1-2x) = -12 < 0.$$

Therefore, we conclude that $x = \frac{1}{2}$ is a maximum point. Hence, the mode of X is $x = \frac{1}{2}$.

4.4.3 Mean

We have defined the mean as the expectation of X. Here, we show how to compute the expectation from the CDF. To simplify the demonstration, let us first assume that $X > 0$.

Lemma 4.1. *Let $X > 0$. Then $\mathbb{E}[X]$ can be computed from F_X as*

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(t))\ dt. \tag{4.17}$$

Proof. The trick is to change the integration order:

$$\begin{aligned}
\int_0^\infty (1 - F_X(t))\ dt &= \int_0^\infty [1 - \mathbb{P}[X \le t]]\ dt = \int_0^\infty \mathbb{P}[X > t]\ dt \\
&= \int_0^\infty \int_t^\infty f_X(x)\ dx\ dt \overset{(a)}{=} \int_0^\infty \int_0^x f_X(x)\ dt\ dx \\
&= \int_0^\infty \int_0^x dt f_X(x)\ dx = \int_0^\infty x f_X(x)\ dx = \mathbb{E}[X].
\end{aligned}$$

Here, step (a) is due to the change of integration order. See **Figure 4.16** for an illustration. □

We draw a picture to illustrate the above lemma. As shown in **Figure 4.17**, the mean of a positive random variable $X > 0$ is equivalent to the area above the CDF.

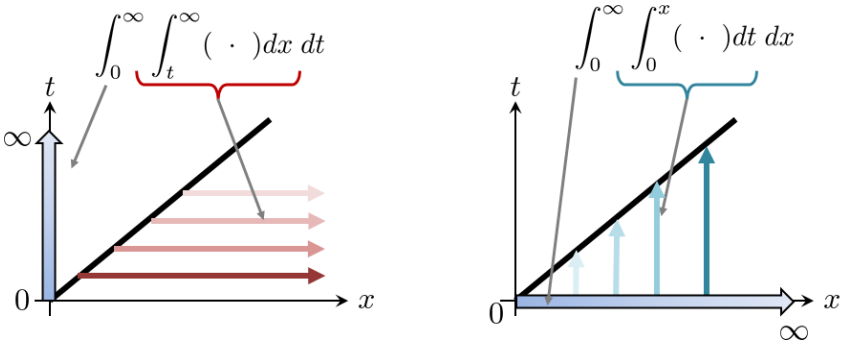


Figure 4.16: The double integration can be evaluated by x then t, or t then x.

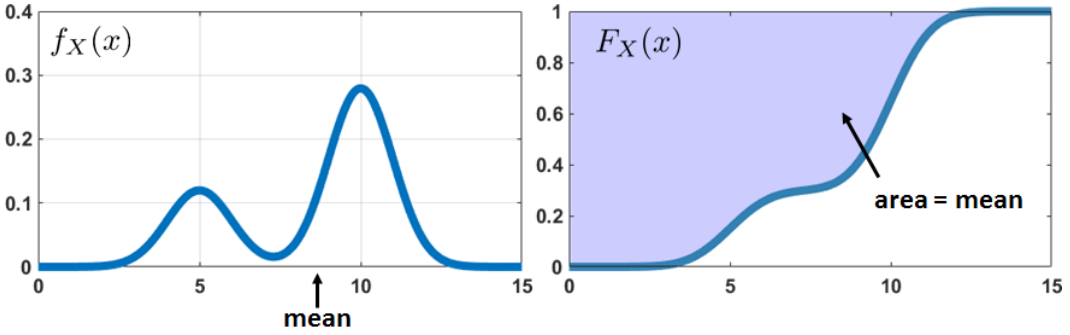


Figure 4.17: The mean of a **positive** random variable $X > 0$ can be calculated by integrating the CDF's complement.

Lemma 4.2. *Let $X < 0$. Then $\mathbb{E}[X]$ can be computed from F_X as*

$$\mathbb{E}[X] = \int_{-\infty}^{0} F_X(t)\ dt. \tag{4.18}$$

Proof. The idea here is also to change the integration order.

$$\int_{-\infty}^{0} F_X(t)\ dt = \int_{-\infty}^{0} \mathbb{P}[X \le t]\ dt = \int_{-\infty}^{0}\int_{-\infty}^{t} f_X(x)\ dx\ dt$$
$$= \int_{-\infty}^{0}\int_{x}^{0} f_X(x)\ dt\ dx = \int_{-\infty}^{0} x f_X(x)\ dx = \mathbb{E}[X].$$

□

Theorem 4.7. *The mean of a random variable X can be computed from the CDF as*

$$\mathbb{E}[X] = \int_{0}^{\infty} (1 - F_X(t))\ dt - \int_{-\infty}^{0} F_X(t)\ dt. \tag{4.19}$$

Proof. For any random variable X, we can partition $X = X^+ - X^-$ where X^+ and X^- are the positive and negative parts, respectively. Then, the above two lemmas will give us

$$\begin{aligned}\mathbb{E}[X] = \mathbb{E}[X^+ - X^-] &= \mathbb{E}[X^+] - \mathbb{E}[X^-] \\ &= \int_0^\infty (1 - F_X(t))\ dt - \int_{-\infty}^0 F_X(t)\ dt.\end{aligned}$$

□

As illustrated in **Figure 4.18**, this equation is equivalent to computing the areas above and below the CDF and taking the difference.

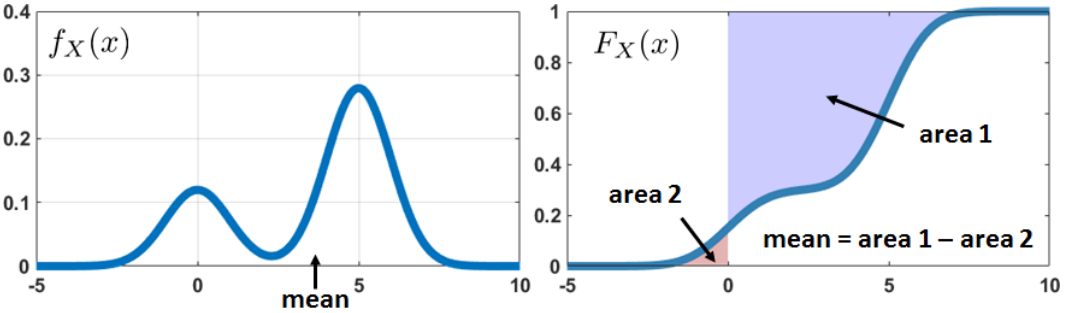


Figure 4.18: The mean of a random variable X can be calculated by computing the area in the CDF.

How to find the mean from the CDF

- A formula is given by Equation (4.20):

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(t))\ dt - \int_{-\infty}^0 F_X(t)\ dt. \tag{4.20}$$

- This result is not commonly used, but the proof technique of switching the integration order is important.

4.5 Uniform and Exponential Random Variables

There are many useful continuous random variables. In this section, we discuss two of them: uniform random variables and exponential random variables. In the next section, we will discuss the Gaussian random variables. Similarly to the way we discussed discrete random variables, we take a generative / synthesis perspective when studying continuous random variables. We assume we have access to the PDF of the random variables so we can derive the theoretical mean and variance. The opposite direction, namely inferring the underlying model parameters from a dataset, will be discussed later.

4.5.1 Uniform random variables

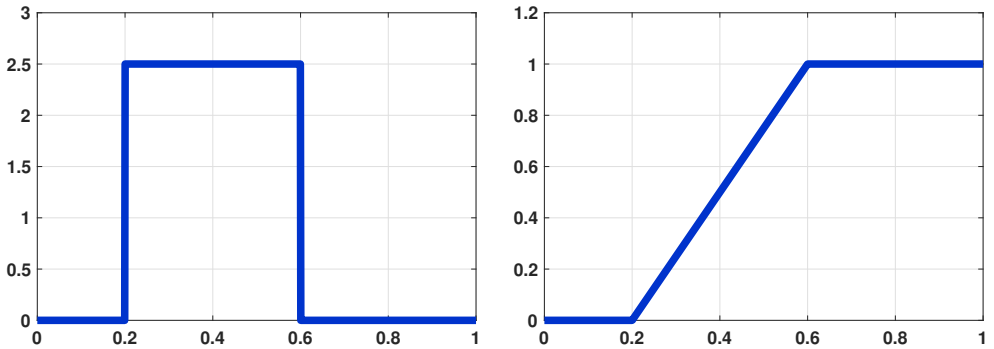
Definition 4.11. *Let X be a continuous* **uniform random variable**. *The PDF of X is*

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise}, \end{cases} \tag{4.21}$$

where $[a, b]$ is the interval on which X is defined. We write

$$X \sim \text{Uniform}(a, b)$$

to mean that X is drawn from a uniform distribution on an interval $[a, b]$.



(a) PDF (b) CDF

Figure 4.19: The PDF and CDF of $X \sim \text{Uniform}(0.2, 0.6)$.

The shape of the PDF of a uniform random variable is shown in **Figure 4.19**. In this figure, we assume that the random variables $X \sim \text{Uniform}(0.2, 0.6)$ are taken from the sample space $\Omega = [0, 1]$. Note that the height of the uniform distribution is greater than 1, since

$$f_X(x) = \begin{cases} \frac{1}{0.6-0.2} = 2.5, & 0.2 \leq x \leq 0.6, \\ 0, & \text{otherwise}. \end{cases}$$

There is nothing wrong with this PDF, because $f_X(x)$ is the probability *per unit length*. If we integrate $f_X(x)$ over any sub-interval between 0.2 and 0.6, we can show that the probability is between 0 and 1.

The CDF of a uniform random variable can be determined by integrating $f_X(x)$:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t)\, dt \\ &= \int_{a}^{x} \frac{1}{b-a}\, dt \\ &= \frac{x-a}{b-a}, \qquad\qquad a \leq x \leq b. \end{aligned}$$

Therefore, the complete CDF is

$$F_X(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

The corresponding CDF for the PDF we showed in **Figure 4.19**(a) is shown in **Figure 4.19**(b). It can be seen that although the height of the PDF exceeds 1, the CDF grows linearly and saturates at 1.

Remark. The uniform distribution can also be defined for discrete random variables. In this case, the probability mass function is given by

$$p_X(k) = \frac{1}{b-a+1}, \quad k = a, a+1, \ldots, b.$$

The presence of "1" in the denominator of the PMF is because k runs from a to b, including the two endpoints.

In MATLAB and Python, generating uniform random numbers can be done by calling commands `unifrnd` (MATLAB), and `stats.uniform.rvs` (Python). For discrete uniform random variables, in MATLAB the command is `unidrnd`, and in Python the command is `stats.randint`.

```
% MATLAB code to generate 1000 uniform random numbers
a = 0; b = 1;
X = unifrnd(a,b,[1000,1]);
hist(X);
```

```
# Python code to generate 1000 uniform random numbers
import scipy.stats as stats
a = 0; b = 1;
X = stats.uniform.rvs(a,b,size=1000)
plt.hist(X);
```

To compute the empirical average and variance of the random numbers in MATLAB we can call the command `mean` and `var`. The corresponding command in Python is `np.mean` and `np.var`. We can also compute the median and mode, as shown below.

```
% MATLAB code to compute empirical mean, var, median, mode
X = unifrnd(a,b,[1000,1]);
M = mean(X);
V = var(X);
Med = median(X);
Mod = mode(X);
```

```
# Python code to compute empirical mean, var, median, mode
X = stats.uniform.rvs(a,b,size=1000)
M = np.mean(X)
V = np.var(X)
```

```
Med = np.median(X)
Mod = stats.mode(X)
```

The mean and variance of a uniform random variable are given by the theorem below.

Theorem 4.8. *If* $X \sim \text{Uniform}(a, b)$, *then*

$$\mathbb{E}[X] = \frac{a+b}{2} \quad \textit{and} \quad \text{Var}[X] = \frac{(b-a)^2}{12}. \tag{4.22}$$

Proof. We have derived these results before. Here is a recap for completeness:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_a^b \frac{x}{b-a}\, dx = \frac{a+b}{2}, \\
\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x)\, dx = \int_a^b \frac{x^2}{b-a}\, dx = \frac{a^2+ab+b^2}{3}, \\
\text{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(b-a)^2}{12}.
\end{aligned}$$

□

The result should be intuitive because it says that the mean is the midpoint of the PDF.

When will we encounter a uniform random variable? Uniform random variables are one of the most elementary continuous random variables. Given a uniform random variable, we can construct any random variable by using an appropriate transformation. We will discuss this technique as part of our discussion about generating random numbers.

In MATLAB, computing the mean and variance of a uniform random variable can be done using the command `unifstat`. The Python coommand is `stats.uniform.stats`.

```
% MATLAB code to compute mean and variance
a = 0; b = 1;
[M,V] = unifstat(a,b)
```

```
# Python code to compute mean and variance
import scipy.stats as stats
a = 0; b = 1;
M, V = stats.uniform.stats(a,b,moments='mv')
```

To evaluate the probability $\mathbb{P}[\ell \leq X \leq u]$ for a uniform random variable, we can call `unifcdf` in MATLAB and

```
% MATLAB code to compute the probability P(0.2 < X < 0.3)
a = 0; b = 1;
F = unifcdf(0.3,a,b) - unifcdf(0.2,a,b)
```

```
# Python code to compute the probability P(0.2 < X < 0.3)
a = 0; b = 1;
F = stats.uniform.cdf(0.3,a,b)-stats.uniform.cdf(0.2,a,b)
```

An alternative is to define an object `rv = stats.uniform`, and call the CDF attribute:

```
# Python code to compute the probability P(0.2 < X < 0.3)
a = 0; b = 1;
rv = stats.uniform(a,b)
F = rv.cdf(0.3)-rv.cdf(0.2)
```

4.5.2 Exponential random variables

Definition 4.12. *Let X be an* **exponential random variable**. *The PDF of X is*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{4.23}$$

where $\lambda > 0$ is a parameter. We write

$$X \sim \text{Exponential}(\lambda)$$

to mean that X is drawn from an exponential distribution of parameter λ.

In this definition, the parameter λ of the exponential random variable determines the rate of decay. A large λ implies a faster decay. The PDF of an exponential random variable is illustrated in **Figure 4.20**. We show two values of λ. Note that the initial value $f_X(0)$ is

$$f_X(0) = \lambda e^{-\lambda 0} = \lambda.$$

Therefore, as long as $\lambda > 1$, $f_X(0)$ will exceed 1.

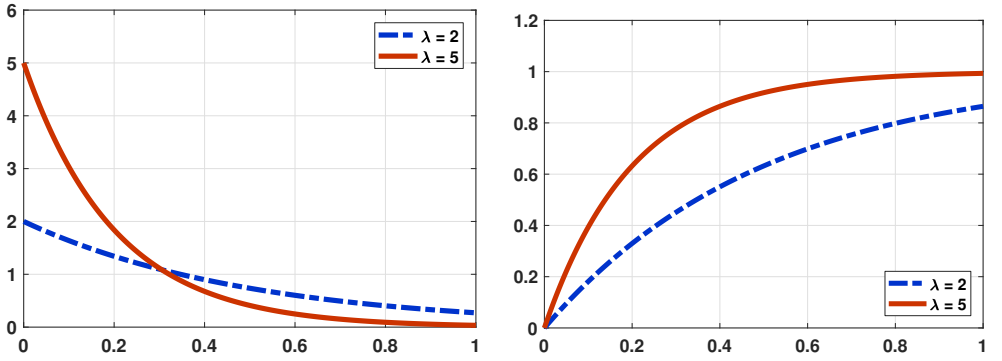
The CDF of an exponential random variable can be determined by

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t)\, dt \\ &= \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}, \qquad x \geq 0. \end{aligned}$$

Therefore, if we consider the entire real line, the CDF is

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

The corresponding CDFs for the PDFs shown in **Figure 4.20**(a) are shown in **Figure 4.20**(b). For larger λ, the PDF $f_X(x)$ decays faster but the CDF $F_X(x)$ increases faster.



(a) PDF (b) CDF

Figure 4.20: (a) The PDF and (c) the CDF of $X \sim$ Exponential(λ).

In MATLAB, the code used to generate **Figure 4.20**(a) is shown below. There are multiple ways of doing this. An alternative way is to call `exppdf`, which will return the same result. In Python, the corresponding command is `stats.expon.pdf`. Note that in Python the parameter λ is specified in `scale` option.

```
% MATLAB code to plot the exponential PDF
lambda1 = 1/2;  lambda2 = 1/5;
x = linspace(0,1,1000);
f1 = pdf('exp',x, lambda1);
f2 = pdf('exp',x, lambda2);
plot(x, f1, 'LineWidth', 4, 'Color', [0 0.2 0.8]); hold on;
plot(x, f2, 'LineWidth', 4, 'Color', [0.8 0.2 0]);
```

```
# Python code to plot the exponential PDF
lambd1 = 1/2
lambd2 = 1/5
x = np.linspace(0,1,1000)
f1 = stats.expon.pdf(x,scale=lambd1)
f2 = stats.expon.pdf(x,scale=lambd2)
plt.plot(x, f1)
plt.plot(x, f2)
```

To plot the CDF, we replace `pdf` by `cdf`. Similarly, in Python we replace `expon.pdf` by `expon.cdf`.

```
% MATLAB code to plot the exponential CDF
F = cdf('exp',x, lambda1);
plot(x, F, 'LineWidth', 4, 'Color', [0 0.2 0.8]);
```

```
# Python code to plot the exponential CDF
F = stats.expon.cdf(x,scale=lambd1)
plt.plot(x, F)
```

Theorem 4.9. *If* $X \sim \text{Exponential}(\lambda)$, *then*

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \textit{and} \quad \text{Var}[X] = \frac{1}{\lambda^2}. \tag{4.24}$$

Proof. We have discussed this proof before. Here is a recap for completeness:

$$\begin{aligned}
\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx &= \int_0^{\infty} \lambda x e^{-\lambda x}\, dx \\
&= -\int_0^{\infty} x de^{-\lambda x} \\
&= -xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x}\, dx = \frac{1}{\lambda},
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx &= \int_0^{\infty} \lambda x^2 e^{-\lambda x}\, dx \\
&= -\int_0^{\infty} x^2 de^{-\lambda x} \\
&= -x^2 e^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} 2x e^{-\lambda x}\, dx \\
&= 0 + \frac{2}{\lambda}\mathbb{E}[X] = \frac{2}{\lambda^2}.
\end{aligned}$$

Thus, $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{\lambda^2}$.

□

Computing the mean and variance of an exponential random variable in MATLAB and Python follows the similar procedures that we described above.

4.5.3 Origin of exponential random variables

Exponential random variables are closely related to Poisson random variables. Recall that the definition of a Poisson random variable is a random variable that describes the number of events that happen in a certain period, e.g., photon arrivals, number of pedestrians, phone calls, etc. We summarize the origin of an exponential random variable as follows.

What is the origin of exponential random variables?

- An exponential random variable is the **interarrival time** between two consecutive Poisson events.
- That is, an exponential random variable is how much time it takes to go from N Poisson counts to $N + 1$ Poisson counts.

An example will clarify this concept. Imagine that you are waiting for a bus, as illustrated in **Figure 4.21**. Passengers arrive at the bus stop with an arrival rate λ per unit time. Thus, for some time t, the average number of people that arrive is λt. Let N be a random

variable denoting the number of people. We assume that N is Poisson with a parameter λt. That is, for any duration t, the probability of observing n people follows the PMF

$$\mathbb{P}[N = n] = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

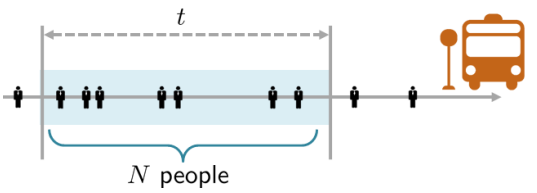


Figure 4.21: For any fixed period of time t, the number of people N is modeled as a Poisson random variable with a parameter λt.

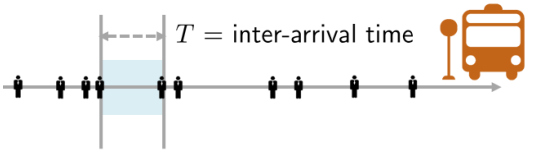


Figure 4.22: The interarrival time T between two consecutive Poisson events is an exponential random variable.

Let T be the interarrival time between two people, by which we mean the time between two consecutive arrivals, as shown in **Figure 4.22**. Note that T is a random variable because T depends on N, which is itself a random variable. To find the PDF of T, we first find the CDF of T. We note that

$$\begin{aligned}\mathbb{P}[T > t] &\overset{(a)}{=} \mathbb{P}[\text{interarrival time } > t] \\ &\overset{(b)}{=} \mathbb{P}[\text{no arrival in } t] \overset{(c)}{=} \mathbb{P}[N = 0] = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}.\end{aligned}$$

In this set of arguments, (a) holds because T is the interarrival time, and (b) holds because interarrival time is between two consecutive arrivals. If the interarrival time is larger than t, there is no arrival during the period. Equality (c) holds because N is the number of passengers.

Since $\mathbb{P}[T > t] = 1 - F_T(t)$, where $F_T(t)$ is the CDF of T, we can show that

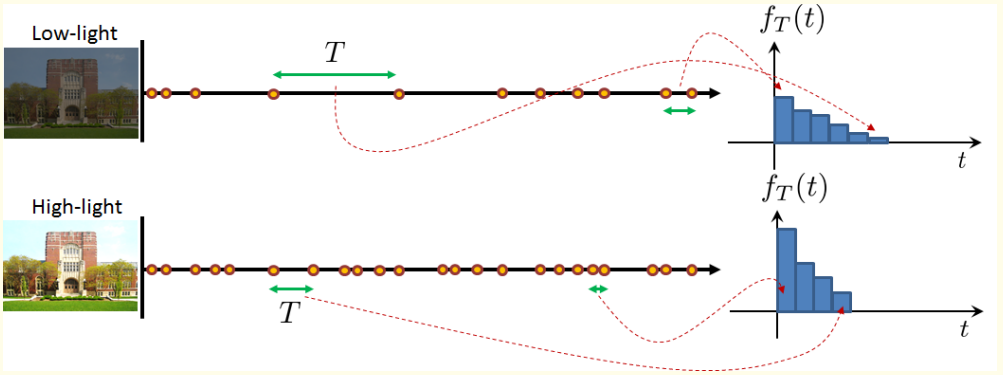
$$\begin{aligned}F_T(t) &= 1 - e^{-\lambda t}, \\ f_T(t) &= \frac{d}{dt} F_T(t) = \lambda e^{-\lambda t}.\end{aligned}$$

Therefore, the interarrival time T follows an exponential distribution.

Since exponential random variables are tightly connected to Poisson random variables, we should expect them to be useful for modeling temporal events. We discuss two examples.

4.5.4 Applications of exponential random variables

Example 4.22. (**Photon arrivals**) Single-photon image sensors are designed to operate in the photon-limited regime. The number-one goal of using these sensors is to count the number of arriving photons precisely. However, for some applications not all single-photon image sensors are used to count photons. Some are used to measure the time between two photon arrivals, such as time-of-flight systems. In this case, we are interested in measuring the time it takes for a pulse to bounce back to the sensor. The more time it takes for a pulse to come back, the greater the distance between the object and the sensor. Other applications utilize the time information. For example, high-dynamic-range imaging can be achieved by recording the time between two photon arrivals because brighter regions have a higher Poisson rate λ and darker regions have a lower λ.



The figure above illustrates an example of high-dynamic-range imaging. When the scene is bright, the large λ will generate more photons. Therefore, the interarrival time between the consecutive photons will be relatively short. If we plot the histogram of the interarrival time, we observe that most of the interarrival time will be concentrated at small values. Dark regions behave in the opposite manner. The interarrival time will typically be much longer. In addition, because there is more variation in the photon arrival times, the histogram will look shorter and wider. Nevertheless, both cases are modeled by the exponential random variable.

Example 4.23. (**Energy-efficient escalator**) Many airports today have installed variable-speed escalators. These escalators change their speeds according to the traffic. If there are no passengers for more than a certain period (say, 60 seconds), the escalator will switch from the full-speed mode to the low-speed mode. For moderately busy escalators, the variable-speed configuration can save energy. The interesting data-science problem is to determine, given a traffic pattern, e.g., the one shown in **Figure 4.23**, whether we can predict the amount of energy savings?

We will not dive into the details of this problem, but we can briefly discuss the principle. Consider a fixed arrival rate λ (say, the average from 07:00 to 08:00). The interarrival time, according to our discussion above, follows an exponential distribution.

So we know that

$$f_T(t) = \lambda e^{-\lambda t}.$$

Suppose that the escalator switches to low-speed mode when the interarrival time exceeds τ. Then we can define a new variable Y to denote the amount of time that the escalator will operate in the low-speed mode. This new variable is

$$Y = \begin{cases} T - \tau, & T > \tau, \\ 0, & T \leq \tau. \end{cases}$$

In other words, if the interarrival time T is more than τ, then the amount of time saved Y takes the value $T - \tau$, but if the interarrival time is less than τ, then there is no saving.

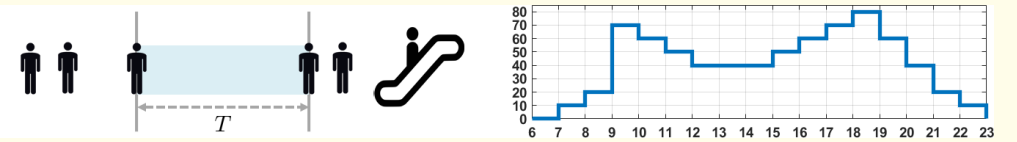


Figure 4.23: The variable-speed escalator problem. [Left] We model the passengers as independent Poisson arrivals. Thus, the interarrival time is exponential. [Right] A hypothetical passenger arrival rate (number of people per minute), from 06:00 to 23:00.

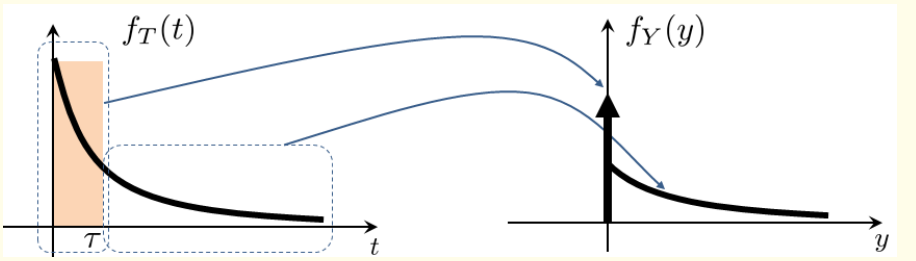


Figure 4.24: The escalator problem requires modeling the cutoff threshold τ such that if $T > \tau$, the savings are $Y = T - \tau$. If $T < \tau$, then $Y = 0$. The left-hand side of the figure shows how the PDF of Y is constructed.

The PDF of Y can be computed according to **Figure 4.24**. There are two parts to the calculation. When $Y = 0$, there is a probability mass such that

$$f_Y(0) = \mathbb{P}[Y = 0] = \int_0^\tau f_T(t)\, dt = \int_0^\tau \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda \tau}.$$

For other values of y, we can show that

$$f_Y(y) = f_T(y + \tau) = \lambda e^{-\lambda(y+\tau)}.$$

Therefore, to summarize, we can show that the PDF of Y is

$$f_Y(y) = \begin{cases} (1 - e^{-\lambda \tau})\delta(y), & y = 0, \\ \lambda e^{-\lambda(y+\tau)}, & y > 0. \end{cases}$$

Consequently, we can compute $\mathbb{E}[Y]$ and $\text{Var}[Y]$ and analyze how these values change for λ (which itself changes with the time of day). Furthermore, we can analyze the amount of savings in terms of dollars. We leave these problems as an exercise.

Closing remark. The photon arrival problem and the escalator problem are two of many examples we can find in which exponential random variables are useful for modeling a problem. We did not go into the details of the problems because each of them requires some additional modeling to address the real practical problem. We encourage you to explore these problems further. Our message is simple: Many problems can be modeled by exponential random variables, most of which are associated with time.

4.6 Gaussian Random Variables

We now discuss *the* most important continuous random variable — the **Gaussian random variable** (also known as the **normal random variable**). We call it the most important random variable because it is widely used in almost all scientific disciplines. Many of us have used Gaussian random variables before, and perhaps its bell shape is the first lesson we learn in statistics. However, there are many mysteries about Gaussian random variables which you may have missed, such as: Where does the Gaussian random variable come from? Why does it take a bell shape? What are the properties of a Gaussian random variable? The objective of this section is to explain everything you need to know about a Gaussian random variable.

4.6.1 Definition of a Gaussian random variable

Definition 4.13. *A* **Gaussian random variable** *is a random variable X such that its PDF is*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{4.25}$$

where (μ, σ^2) are parameters of the distribution. We write

$$X \sim \textit{Gaussian}(\mu, \sigma^2) \qquad \textit{or} \qquad X \sim \mathcal{N}(\mu, \sigma^2)$$

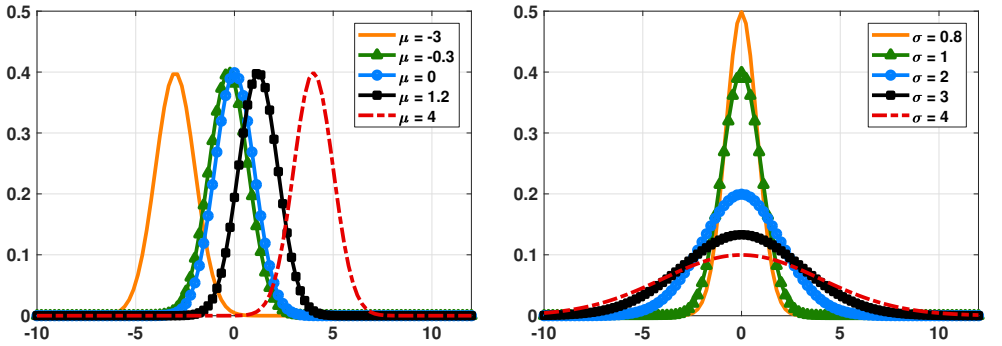
to say that X is drawn from a Gaussian distribution of parameter (μ, σ^2).

Gaussian random variables have two parameters (μ, σ^2). It is noteworthy that the mean is μ and the variance is σ^2 — these two parameters are exactly the first moment and the second central moment of the random variable. Most other random variables do not have this property.

Note that a Gaussian random variable is positive from $-\infty$ to ∞. Thus, $f_X(x)$ has a non-zero value for any x, even though the value may be extremely small. A Gaussian random variable is also symmetric about μ. If $\mu = 0$, then $f_X(x)$ is an even function.

The shape of the Gaussian is illustrated in **Figure 4.25**. When we fix the variance and change the mean, the PDF of the Gaussian moves left or right depending on the sign of the mean. When we fix the mean and change the variance, the PDF of the Gaussian changes

its width. Since any PDF should integrate to unity, a wider Gaussian means that the PDF is shorter. Note also that if σ is very small, it is possible that $f_X(x) > 1$ although the integration over Ω will still be 1.



μ changes, $\sigma = 1$ $\mu = 0$, σ changes

Figure 4.25: A Gaussian random variable with different μ and σ.

On a computer, plotting the Gaussian PDF can be done by calling the function `pdf('norm',x)` in MATLAB, and `stats.norm.pdf` in Python.

```
% MATLAB to generate a Gaussian PDF
x     = linspace(-10,10,1000);
mu    = 0; sigma = 1;
f     = pdf('norm',x,mu,sigma);
plot(x, f);
```

```
# Python to generate a Gaussian PDF
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
x  = np.linspace(-10,10,1000)
mu = 0; sigma = 1;
f  = stats.norm.pdf(x,mu,sigma)
plt.plot(x,f)
```

Our next result concerns the mean and variance of a Gaussian random variable. You may wonder why we need this theorem when we already know that μ is the mean and σ^2 is the variance. The answer is that we have not proven these two facts.

Theorem 4.10. *If $X \sim$ Gaussian(μ, σ^2), then*

$$\mathbb{E}[X] = \mu, \quad \textit{and} \quad \mathrm{Var}[X] = \sigma^2. \tag{4.26}$$

Proof. The expectation can be derived via substitution:

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\
&\overset{(a)}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (y+\mu) e^{-\frac{y^2}{2\sigma^2}}\, dy \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} y e^{-\frac{y^2}{2\sigma^2}}\, dy + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \mu e^{-\frac{y^2}{2\sigma^2}}\, dy \\
&\overset{(b)}{=} 0 + \mu \left(\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma^2}}\, dy \right) \\
&\overset{(c)}{=} \mu,
\end{aligned}$$

where in (a) we substitute $y = x - \mu$, in (b) we use the fact that the first integrand is odd so that the integration is 0, and in (c) we observe that integration over the entire sample space of the PDF yields 1.

The variance is also derived by substitution.

$$\begin{aligned}
\mathrm{Var}[X] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\
&\overset{(a)}{=} \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2}}\, dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(-y e^{-\frac{y^2}{2}} \Big|_{-\infty}^{\infty} \right) + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\, dy \\
&= 0 + \sigma^2 \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\, dy \right) \\
&= \sigma^2,
\end{aligned}$$

where in (a) we substitute $y = (x - \mu)/\sigma$.

4.6.2 Standard Gaussian

We need to evaluate the probability $\mathbb{P}[a \le X \le b]$ of a Gaussian random variable X in many practical situations. This involves the integration of the Gaussian PDF, i.e., determining the CDF. Unfortunately, there is no closed-form expression of $\mathbb{P}[a \le X \le b]$ in terms of (μ, σ^2). This leads to what we call the standard Gaussian.

Definition 4.14. *The* **standard Gaussian** *(or standard normal) random variable X has a PDF*

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{4.27}$$

That is, $X \sim \mathcal{N}(0,1)$ is a Gaussian with $\mu = 0$ and $\sigma^2 = 1$.

The CDF of the standard Gaussian can be determined by integrating the PDF. We have a special notation for this CDF. **Figure 4.26** illustrates the idea.

Definition 4.15. *The* **CDF** *of the standard Gaussian is defined as the* $\Phi(\cdot)$ *function*

$$\Phi(x) \overset{\text{def}}{=} F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt. \tag{4.28}$$

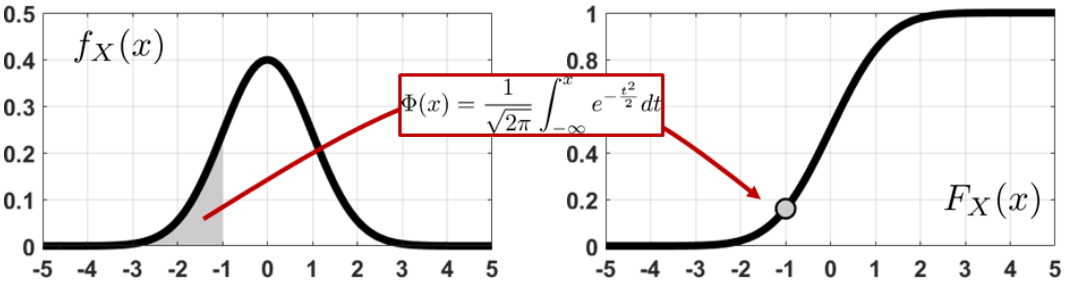


Figure 4.26: Definition of the CDF of the standard Gaussian $\Phi(x)$.

```
% MATLAB code to generate standard Gaussian PDF and CDF
x = linspace(-5,5,1000);
f = normpdf(x,0,1);
F = normcdf(x,0,1);
figure; plot(x, f);
figure; plot(x, F);
```

```
# Python code to generate standard Gaussian PDF and CDF
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
x = np.linspace(-10,10,1000)
f = stats.norm.pdf(x)
F = stats.norm.cdf(x)
plt.plot(x,f); plt.show()
plt.plot(x,F); plt.show()
```

The standard Gaussian's CDF is related to a so-called **error function** defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2}\, dt. \tag{4.29}$$

It is easy to link $\Phi(x)$ with $\text{erf}(x)$:

$$\Phi(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right], \qquad \text{and} \qquad \text{erf}(x) = 2\Phi(x\sqrt{2}) - 1.$$

With the standard Gaussian CDF, we can define the CDF of an arbitrary Gaussian.

Theorem 4.11 (**CDF of an arbitrary Gaussian**). *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then*

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{4.30}$$

Proof. We start by expressing $F_X(x)$:

$$\begin{aligned} F_X(x) &= \mathbb{P}[X \leq x] \\ &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt. \end{aligned}$$

Substituting $y = \frac{t-\mu}{\sigma}$, and using the definition of standard Gaussian, we have

$$\begin{aligned} \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right). \end{aligned}$$

□

If you would like to verify this on a computer, you can try the following code.

```
% MATLAB code to verify standardized Gaussian
x = linspace(-5,5,1000);
mu = 3; sigma = 2;
f1 = normpdf((x-mu)/sigma,0,1); % standardized
f2 = normpdf(x, mu, sigma);     % raw
```

```
# Python code to verify standardized Gaussian
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
x = np.linspace(-5,5,1000)
mu = 3; sigma = 2;
f1 = stats.norm.pdf((x-mu)/sigma,0,1) # standardized
f2 = stats.norm.cdf(x,mu,sigma)       # raw
```

An immediate consequence of this result is that

$$\mathbb{P}[a < X \leq b] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{4.31}$$

To see this, note that

$$\begin{aligned} \mathbb{P}[a < X \leq b] &= \mathbb{P}[X \leq b] - \mathbb{P}[X \leq a] \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

The inequality signs of the two end points are not important. That is, the statement also holds for $\mathbb{P}[a \leq X \leq b]$ or $\mathbb{P}[a < X < b]$, because X is a continuous random variable at every x. Thus, $\mathbb{P}[X = a] = \mathbb{P}[X = b] = 0$ for any a and b. Besides this, Φ has several properties of interest. See if you can prove these:

Corollary 4.1. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then the following results hold:*

- $\Phi(y) = 1 - \Phi(-y)$.
- $\mathbb{P}[X \geq b] = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$.
- $\mathbb{P}[|X| \geq b] = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right) + \Phi\left(\frac{-b-\mu}{\sigma}\right)$.

4.6.3 Skewness and kurtosis

In modern data analysis we are sometimes interested in high-order moments. Here we consider two useful quantities: **skewness** and **kurtosis**.

Definition 4.16. *For a random variable X with PDF $f_X(x)$, define the following* **central moments** *as*

$$\begin{aligned}
\textit{mean} &= \mathbb{E}[X] \stackrel{\text{def}}{=} \mu, \\
\textit{variance} &= \mathbb{E}\left[(X-\mu)^2\right] \stackrel{\text{def}}{=} \sigma^2, \\
\textit{skewness} &= \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \stackrel{\text{def}}{=} \gamma, \\
\textit{kurtosis} &= \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] \stackrel{\text{def}}{=} \kappa, \qquad \textit{excess kurtosis} \stackrel{\text{def}}{=} \kappa - 3.
\end{aligned}$$

As you can see from the definitions above, skewness is the third central moment, whereas kurtosis is the fourth central moment. Both skewness and kurtosis can be regarded as "deviations" from a standard Gaussian —not in terms of mean and variance but in terms of shape.

Skewness measures the **asymmetry** of the distribution. **Figure 4.27** shows three different distributions: one with left skewness, one with right skewness, and one symmetric. The skewness of a curve is

- Skewed towards left: positive
- Skewed towards right: negative
- Symmetric: zero

What is skewness?

- $\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$.
- Measures the **asymmetry** of the distribution.
- Gaussian has skewness 0.

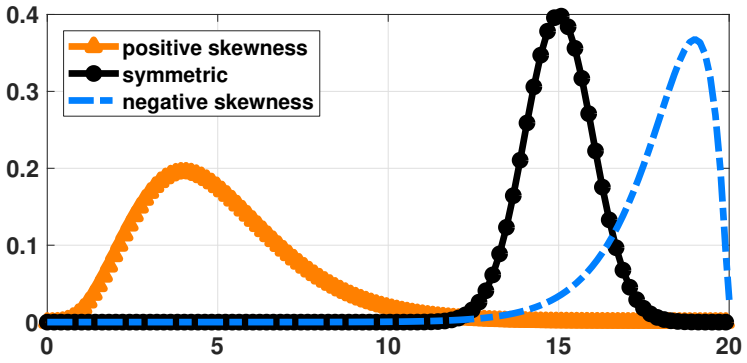


Figure 4.27: Skewness of a distribution measures the asymmetry of the distribution. In this example the skewnesses are: orange = 0.8943, black = 0, blue = -1.414.

Kurtosis measures how **heavy-tailed** the distribution is. There are two forms of kurtosis: one is the standard kurtosis, which is the fourth central moment, and the other is the **excess** kurtosis, which is $\kappa_{\text{excess}} = \kappa - 3$. The constant 3 comes from the kurtosis of a standard Gaussian. Excess kurtosis is more widely used in data analysis. The interpretation of kurtosis is the comparison to a Gaussian. If the kurtosis is positive, the distribution has a tail that decays faster than a Gaussian. If the kurtosis is negative, the distribution has a tail that decays more slowly than a Gaussian. **Figure 4.28** illustrates the (excess) kurtosis of three different distributions.

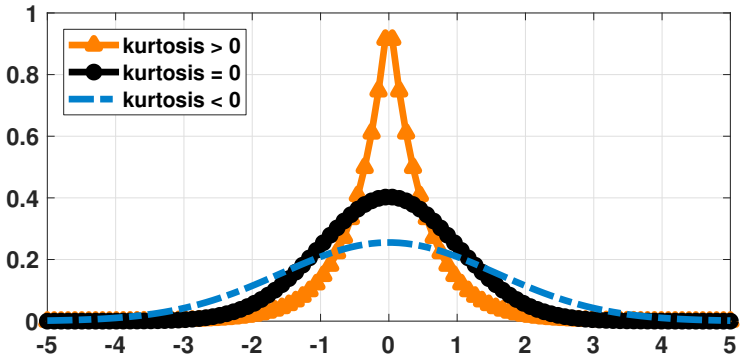


Figure 4.28: Kurtosis of a distribution measures how heavy-tailed the distribution is. In this example, the (excess) kurtoses are: orange = 2.8567, black = 0, blue = -0.1242.

What is kurtosis?

- $\kappa = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$.
- Measures how **heavy-tailed** the distribution is. Gaussian has kurtosis 3.
- Some statisticians prefer **excess kurtosis** $\kappa - 3$, so that Gaussian has excess kurtosis 0.

Random variable	Mean μ	Variance σ^2	Skewness γ	Excess kurtosis $\kappa - 3$
Bernoulli	p	$p(1-p)$	$\frac{1-2p}{\sqrt{p(1-p)}}$	$\frac{1}{1-p} + \frac{1}{p} - 6$
Binomial	np	$np(1-p)$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{6p^2-6p+1}{np(1-p)}$
Geometric	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{2-p}{\sqrt{1-p}}$	$\frac{p^2-6p+6}{1-p}$
Poisson	λ	λ	$\frac{1}{\sqrt{\lambda}}$	$\frac{1}{\lambda}$
Uniform	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	0	$-\frac{6}{5}$
Exponential	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	2	6
Gaussian	μ	σ^2	0	0

Table 4.1: The first few moments of commonly used random variables.

On a computer, computing the **empirical** skewness and kurtosis is done by built-in commands. Their implementations are based on the finite-sample calculations

$$\gamma \approx \frac{1}{N}\sum_{n=1}^{N}\left(\frac{X_n - \mu}{\sigma}\right)^3,$$

$$\kappa \approx \frac{1}{N}\sum_{n=1}^{N}\left(\frac{X_n - \mu}{\sigma}\right)^4.$$

The MATLAB and Python built-in commands are shown below, using a gamma distribution as an example.

```
% MATLAB code to compute skewness and kurtosis
X = random('gamma',3,5,[10000,1]);
s = skewness(X);
k = kurtosis(X);
```

```
# Python code to compute skewness and kurtosis
import scipy.stats as stats
X = stats.gamma.rvs(3,5,size=10000)
s = stats.skew(X)
k = stats.kurtosis(X)
```

Example 4.24. To further illustrate the behavior of skewness and kurtosis, we consider an example using the gamma random variable X. The PDF of X is given by the equation

$$f_X(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}, \tag{4.32}$$

where $\Gamma(\cdot)$ is known as the gamma function. If k is an integer, the gamma function is

just the factorial: $\Gamma(k) = (k-1)!$. A gamma random variable is parametrized by two parameters (k, θ). As k increases or decreases, the shape of the PDF will change. For example, when $k = 1$, the distribution is simplified to an exponential distribution.

Without going through the (tedious) integration, we can show that the skewness and the (excess) kurtosis of Gamma(k, θ) are

$$\text{skewness} = \frac{2}{\sqrt{k}},$$
$$\text{(excess) kurtosis} = \frac{6}{k}.$$

As we can see from these results, the skewness and kurtosis diminish as k grows. This can be confirmed from the PDF of Gamma(k, θ) as shown in **Figure 4.29**.

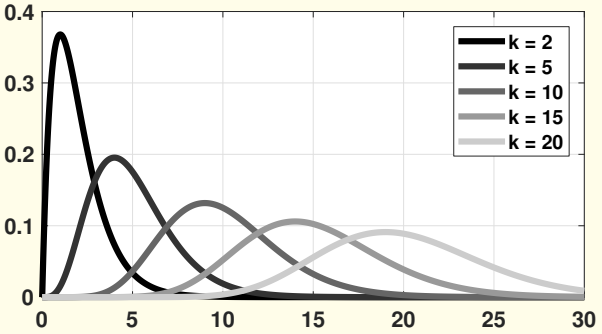
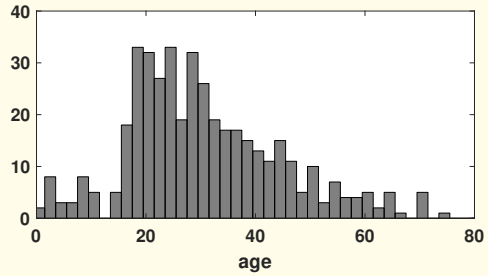
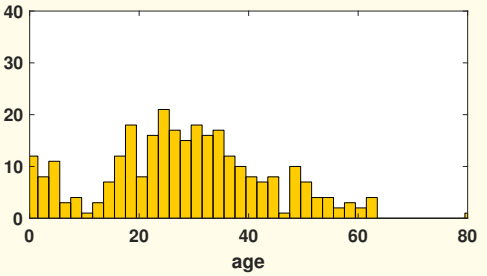


Figure 4.29: The PDF of a gamma distribution Gamma(k, θ), where $\theta = 1$. The skewness and the kurtosis are decaying to zero.

Example 4.25. Let us look at a real example. On April 15, 1912, RMS *Titanic* sank after hitting an iceberg. The disaster killed 1502 out of 2224 passengers and crew. A hundred years later, we want to analyze the data. At `https://www.kaggle.com/c/titanic/` there is a dataset collecting the identities, age, gender, etc., of the passengers. We partition the dataset into two: one for those who died and the other one for those who survived. We plot the histograms of the ages of the two groups and compute several statistics of the dataset. **Figure 4.30** shows the two datasets.



Group 1 (died)



Group 2 (survived)

Figure 4.30: The Titanic dataset `https://www.kaggle.com/c/titanic/`.

Statistics	Group 1 (Died)	Group 2 (Survived)
Mean	30.6262	28.3437
Standard Deviation	14.1721	14.9510
Skewness	0.5835	0.1795
Excess Kurtosis	0.2652	−0.0772

Note that the two groups of people have very similar means and standard deviations. In other words, if we only compare the mean and standard deviation, it is nearly impossible to differentiate the two groups. However, the skewness and kurtosis provide more information related to the shape of the histograms. For example, Group 1 has more positive skewness, whereas Group 2 is almost symmetrical. One interpretation is that more young people offered lifeboats to children and older people. The kurtosis of Group 1 is slightly positive, whereas that of Group 2 is slightly negative. Therefore, high-order moments can sometimes be useful for data analysis.

4.6.4 Origin of Gaussian random variables

The Gaussian random variable has a long history. Here, we provide one perspective on why Gaussian random variables are so useful. We give some intuitive arguments but leave the formal mathematical treatment for later when we introduce the Central Limit Theorem.

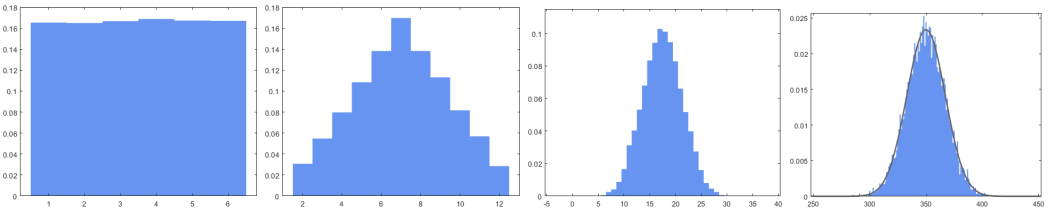
Let's begin with a numerical experiment. Consider throwing a fair die. We know that this will give us a (discrete) uniform random variable X. If we repeat the experiment many times we can plot the histogram, and it will return us a plot of 6 impulses with equal height, as shown in **Figure 4.31**(a).

Now, suppose we throw two dice. Call them X_1 and X_2, and let $Z = X_1 + X_2$, i.e., the sum of two dice. We want to find the distribution of Z. To do so, we first list out all the possible outcomes in the sample space; this gives us $\{(1,1),(1,2),\ldots,(6,6)\}$. We then sum the numbers, which gives us a list of states of Z: $\{2,3,4,\ldots,12\}$. The probability of getting these states is shown in **Figure 4.31**(b), which has a triangular shape. The triangular shape makes sense because to get the state "2", we must have the pair $(1,1)$, which is quite unlikely. However, if we want to get the state 7, it would be much easier to get a pair, e.g., $(6,1),(5,2),(4,3),(3,4),(2,5),(1,6)$ would all do the job.

Now, what will happen if we throw 5 dice and consider $Z = X_1+X_2+\cdots+X_5$? It turns out that the distribution will continue to evolve and give something like **Figure 4.31**(c). This is starting to approximate a bell shape. Finally, if we throw 100 dice and consider $Z = X_1 + X_2 + \cdots + X_{100}$, the distribution will look like **Figure 4.31**(d). The shape is becoming a Gaussian! This numerical example demonstrates a fascinating phenomenon: As we sum more random variables, the distribution of the sum will eventually converge to a Gaussian.

If you are curious about how we plot the above figures, the following MATLAB and Python code can be useful.

```
% MATLAB code to show the histogram of Z = X1+X2+X3
N  = 10000;
X1 = randi(6,1,N);
X2 = randi(6,1,N);
```



(a) X_1 (b) $X_1 + X_2$ (c) $X_1 + \cdots + X_5$ (d) $X_1 + \cdots + X_{100}$

Figure 4.31: When adding uniform random variables, the overall distribution approaches a Gaussian as the number of summed variables increase.

```
X3 = randi(6,1,N);
Z  = X1 + X2 + X3;
histogram(Z, 2.5:18.5);
```

```
# Python code to show the histogram of Z = X1+X2+X3
import numpy as np
import matplotlib.pyplot as plt
N = 10000
X1 = np.random.randint(1,6,size=N)
X2 = np.random.randint(1,6,size=N)
X3 = np.random.randint(1,6,size=N)
Z  = X1 + X2 + X3
plt.hist(Z,bins=np.arange(2.5,18.5))
```

Can we provide a more formal description of this? Yes, but we need some new mathematical tools that we have not yet developed. So, for the time being, we will outline the flow of the arguments and leave the technical details to a later chapter. Suppose we have two independent random variables with identical distributions, e.g., X_1 and X_2, where both are uniform. This gives us PDFs $f_{X_1}(x)$ and $f_{X_2}(x)$ that are two identical rectangular functions. By what operation can we combine these two rectangular functions and create a triangle function? The key lies in the concept of **convolution**. If you convolve two rectangle functions, you will get a triangle function. Here we define the convolution of f_X as

$$(f_X * f_X)(x) = \int_{-\infty}^{\infty} f_X(\tau) f_X(x - \tau)\, d\tau.$$

In fact, for any pair of random variables X_1 and X_2 (not necessarily uniform random variables), the sum $Z = X_1 + X_2$ will have a PDF given by the convolution of the two PDFs. We have not yet proven this, but if you trust what we are saying, we can effectively generalize this argument to many random variables. If we have N random variables, then the sum $Z = X_1 + X_2 + \cdots + X_N$ will have a PDF that is the result of N convolutions of all the individual PDFs.

What is the PDF of $X + Y$?

- Summing $X + Y$ is equivalent to convolving the PDFs $f_X * f_Y$.

- If you sum many random variables, you convolve all their PDFs.

How do we analyze these convolutions? We need a second set of tools related to Fourier transforms. The Fourier transform of a PDF is known as the *characteristic function*, which we will discuss later, but the name is not important now. What matters is the important property of the Fourier transform, that a convolution in the original space is multiplication in the Fourier space. That is,

$$\mathcal{F}\{(f_X * f_X * \cdots * f_X)\} = \mathcal{F}\{f_X\} \cdot \mathcal{F}\{f_X\} \cdot \cdots \cdot \mathcal{F}\{f_X\}.$$

Multiplication in the Fourier space is much easier to analyze. In particular, for independent and identically distributed random variables, the multiplication will easily translate to addition in the exponent. Then, by truncating the exponent to the second order, we can show that the limiting object in the Fourier space is approaching a Gaussian. Finally, since the inverse Fourier transform of a Gaussian remains a Gaussian, we have shown that the infinite convolution will give us a Gaussian.

Here is some numerical evidence for what we have just described. Recall that the Fourier transform of a rectangle function is the sinc function. Therefore, if we have an infinite convolution of rectangular functions, equivalently, we have an infinite product of sinc functions in the Fourier space. Multiplying sinc functions is reasonably easy. See **Figure 4.32** for the first three sincs. It is evident that with just three sinc functions, the shape closely approximates a Gaussian.

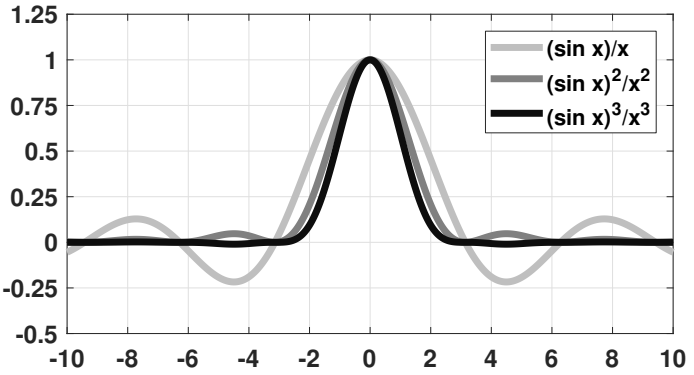


Figure 4.32: Convolving the PDF of a uniform distribution is equivalent to multiplying their Fourier transforms in the Fourier space. As the number of convolutions grows, the product is gradually becoming Gaussian.

How about distributions that are not rectangular? We invite you to numerically visualize the effect when you convolve the function many times. You will see that as the number of convolutions grows, the resulting function will become more and more like a Gaussian. Regardless of what the input random variables are, as long as you add them, the sum will have a distribution that looks like a Gaussian:

$$X_1 + X_2 + \cdots + X_N \rightsquigarrow \text{Gaussian}.$$

We use the notation $\rightsquigarrow$ to emphasize that the convergence is not the usual form of convergence. We will make this precise later.

The implication of this line of discussion is important. Regardless of the underlying true physical process, if we are only interested in the sum (or average), the distribution will be more or less Gaussian. In most engineering problems, we are looking at the sum or average. For example, when generating an image using an image sensor, the sensor will add a certain amount of read noise. Read noise is caused by the random fluctuation of the electrons in the transistors due to thermal distortions. For high-photon-flux situations, we are typically interested in the average read noise rather than the electron-level read noise. Thus Gaussian random variables become a reasonable model for that. In other applications, such as imaging through a turbulent medium, the random phase distortions (which alter the phase of the wavefront) can also be modeled as a Gaussian random variable. Here is the summary of the origin of a Gaussian random variable:

What is the origin of Gaussian?

- When we **sum** many independent random variables, the resulting random variable is a Gaussian.
- This is known as the **Central Limit Theorem**. The theorem applies to *any* random variable.
- Summing random variables is equivalent to **convolving** the PDFs. Convolving PDFs infinitely many times yields the bell shape.

4.7 Functions of Random Variables

One common question we encounter in practice is the transformation of random variables. The question can be summarized as follows: Given a random variable X with PDF $f_X(x)$ and CDF $F_X(x)$, and supposing that $Y = g(X)$ for some function g, what are $f_Y(y)$ and $F_Y(y)$? This is a prevalent question. For example, we measure the voltage V, and we want to analyze the power $P = V^2/R$. This involves taking the square of a random variable. Another example: We know the distribution of the phase Θ, but we want to analyze the signal $\cos(\omega t + \Theta)$. This involves a cosine transformation. How do we convert one variable to another? Answering this question is the goal of this section.

4.7.1 General principle

We will first outline the general principle for tackling this type of problem. In the following subsection, we will give a few concrete examples.

Suppose we are given a random variable X with PDF $f_X(x)$ and CDF $F_X(x)$. Let $Y = g(X)$ for some known and fixed function g. For simplicity, we assume that g is monotonically

increasing. In this case, the CDF of Y can be determined as follows.

$$\begin{aligned} F_Y(y) &\overset{(a)}{=} \mathbb{P}[Y \leq y] \overset{(b)}{=} \mathbb{P}[g(X) \leq y] \\ &\overset{(c)}{=} \mathbb{P}[X \leq g^{-1}(y)] \\ &\overset{(d)}{=} F_X(g^{-1}(y)). \end{aligned}$$

This sequence of steps is not difficult to understand. Step (a) is the definition of CDF. Step (b) substitutes $g(X)$ for Y. Step (c) uses the fact that since g is invertible, we can apply the inverse of g to both sides of $g(X) \leq y$ to yield $X \leq g^{-1}(y)$. Step (d) is the definition of the CDF, but this time applied to $\mathbb{P}[X \leq \clubsuit] = F_X(\clubsuit)$, for some $\clubsuit$.

It will be useful to visualize the situation in **Figure 4.33**. Here, we consider a uniformly distributed X so that the CDF $F_X(x)$ is a straight line. According to F_X, any samples drawn according to F_X are equally likely, as illustrated by the yellow dots on the x-axis. As we transform the X's through $Y = g(X)$, we increase/decrease the spacing between two samples. Therefore, some samples become more concentrated while some become less concentrated. The distribution of these transformed samples (the yellow dots on the y-axis) forms a new CDF $F_Y(y)$. The result $F_Y(y) = F_X(g^{-1}(y))$ holds when we look at Y. The samples are traveling with g^{-1} in order to go back to F_X. Therefore, we need g^{-1} in the formula.

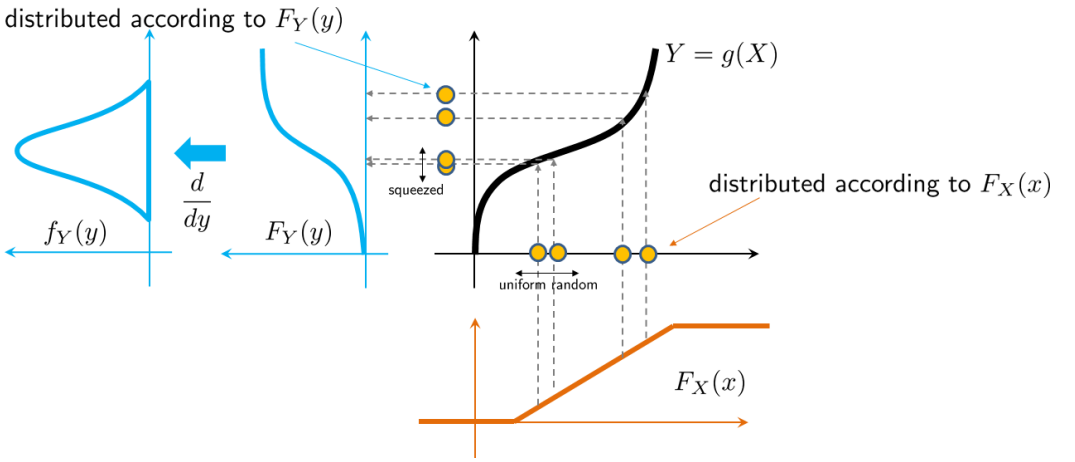


Figure 4.33: When transforming a random variable X to $Y = g(X)$, the distributions are defined according to the spacing between samples. In this figure, a uniformly distributed X will become squeezed by some parts of g and widened in other parts of g.

Why should we use the CDF and not the PDF in **Figure 4.33**? The advantage of the CDF is that it is an increasing function. Therefore, no matter what the function g is, the input and the output functions will still be increasing. If we use the PDF, then the non-monotonic behavior of the PDF will interact with another nonlinear function g. It becomes much harder to decouple the two.

We can carry out the integrations to determine $F_X(g^{-1}(y))$. It can be shown that

$$F_X(g^{-1}(y)) = \int_{-\infty}^{g^{-1}(y)} f_X(x')\, dx', \tag{4.33}$$

and hence, by the fundamental theorem of calculus, we have

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = \frac{d}{dy}\int_{-\infty}^{g^{-1}(y)} f_X(x')\,dx' = \left(\frac{d\,g^{-1}(y)}{dy}\right)\cdot f_X(g^{-1}(y)), \tag{4.34}$$

where the last step is due to the chain rule. Based on this line of reasoning we can summarize a "recipe" for this problem.

How to find the PDF of $Y = g(X)$

- Step 1: Find the CDF $F_Y(y)$, which is $F_Y(y) = F_X(g^{-1}(y))$.
- Step 2: Find the PDF $f_Y(y)$, which is $f_Y(y) = \left(\frac{d\,g^{-1}(y)}{dy}\right)\cdot f_X(g^{-1}(y))$.

This recipe works when g is a one-to-one mapping. If g is not one-to-one, e.g., $g(x) = x^2$ implies $g^{-1}(y) = \pm\sqrt{y}$, then we will have some issues with the above two steps. When this happens, then instead of writing $X \le g^{-1}(y)$ we need to determine the set $\{x \mid g(x) \le y\}$.

4.7.2 Examples

Example 4.26. (Linear transform) Let X be a random variable with PDF $f_X(x)$ and CDF $F_X(x)$. Let $Y = 2X + 3$. Find $f_Y(y)$ and $F_Y(y)$. Express the answers in terms of $f_X(x)$ and $F_X(x)$.

Solution. We first note that

$$\begin{aligned} F_Y(y) &= \mathbb{P}[Y \le y] \\ &= \mathbb{P}[2X + 3 \le y] \\ &= \mathbb{P}\left[X \le \frac{y-3}{2}\right] = F_X\left(\frac{y-3}{2}\right). \end{aligned}$$

Therefore, the PDF is

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}F_Y(y) \\ &= \frac{d}{dy}F_X\left(\frac{y-3}{2}\right) \\ &= F_X'\left(\frac{y-3}{2}\right)\frac{d}{dy}\left(\frac{y-3}{2}\right) = \frac{1}{2}f_X\left(\frac{y-3}{2}\right). \end{aligned}$$

Follow-Up. (Linear transformation of a Gaussian random variable).Suppose X is a Gaussian random variable with zero mean and unit variance, and let $Y = aX + b$. Then the CDF

and PDF of Y are respectively

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right) = \Phi\left(\frac{y-b}{a}\right),$$
$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi}a} e^{-\frac{(y-b)^2}{2a^2}}.$$

Follow-Up. (Linear transformation of an exponential random variable). Suppose X is an exponential random variable with parameter λ, and let $Y = aX + b$. Then the CDF and PDF of Y are respectively

$$\begin{aligned} F_Y(y) &= F_X\left(\frac{y-b}{a}\right) \\ &= 1 - e^{-\frac{\lambda}{a}(y-b)}, \qquad y \geq b, \\ f_Y(y) &= \frac{1}{a} f_X\left(\frac{y-b}{a}\right) \\ &= \frac{\lambda}{a} e^{-\frac{\lambda}{a}(y-b)}, \qquad y \geq b. \end{aligned}$$

Example 4.27. Let X be a random variable with PDF $f_X(x)$ and CDF $F_X(x)$. Supposing that $Y = X^2$, find $f_Y(y)$ and $F_Y(y)$. Express the answers in terms of $f_X(x)$ and $F_X(x)$.

Solution. We note that

$$\begin{aligned} F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[X^2 \leq y] &= \mathbb{P}[-\sqrt{y} \leq X \leq \sqrt{y}] \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

Therefore, the PDF is

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy}\left(F_X(\sqrt{y}) - F_X(-\sqrt{y})\right) \\ &= F_X'(\sqrt{y}) \frac{d}{dy}\sqrt{y} - F_X'(-\sqrt{y}) \frac{d}{dy}\left(-\sqrt{y}\right) \\ &= \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right). \end{aligned}$$

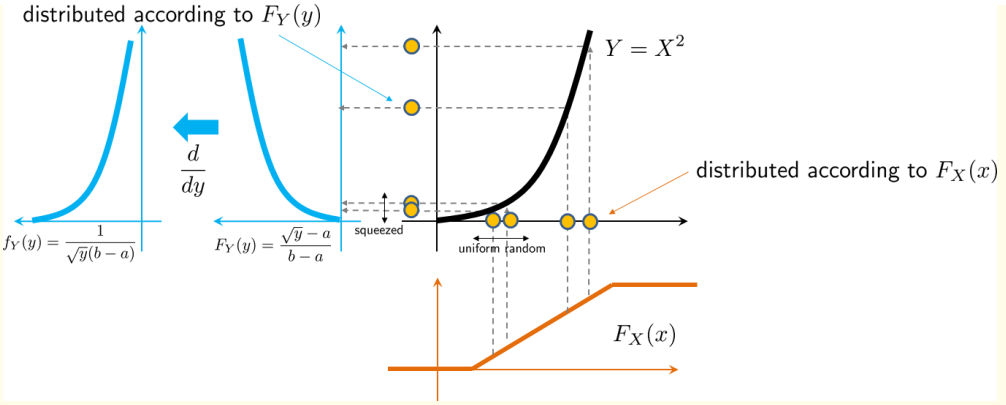


Figure 4.34: When transforming a random variable X to $Y = X^2$, the CDF becomes $F_Y(y) = \frac{\sqrt{y}-a}{b-a}$ and the PDF becomes $f_Y(y) = \frac{1}{\sqrt{y}(b-a)}$.

Follow Up. (Square of a uniform random variable) Suppose X is a uniform random variable in $[a, b]$ (assume $a > 0$), and let $Y = X^2$. Then the CDF and PDF of Y are respectively

$$F_Y(y) = \frac{\sqrt{y} - a}{b - a}, \qquad a^2 \le y \le b^2,$$
$$f_Y(y) = \frac{1}{\sqrt{y}(b-a)}, \qquad a^2 \le y \le b^2.$$

Example 4.28. Let $X \sim \text{Uniform}(0, 2\pi)$. Suppose $Y = \cos X$. Find $f_Y(y)$ and $F_Y(y)$.

Solution. First, we need to find the CDF of X. This can be done by noting that

$$F_X(x) = \int_{-\infty}^{x} f_X(x')\, dx' = \int_0^x \frac{1}{2\pi}\, dx' = \frac{x}{2\pi}.$$

Thus, the CDF of Y is

$$\begin{aligned} F_Y(y) &= \mathbb{P}[Y \le y] = \mathbb{P}[\cos X \le y] \\ &= \mathbb{P}[\cos^{-1} y \le X \le 2\pi - \cos^{-1} y] \\ &= F_X(2\pi - \cos^{-1} y) - F_X(\cos^{-1} y) \\ &= 1 - \frac{\cos^{-1} y}{\pi}. \end{aligned}$$

The PDF of Y is

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy}\left(1 - \frac{\cos^{-1} y}{\pi}\right) \\ &= \frac{1}{\pi\sqrt{1-y^2}}, \end{aligned}$$

where we used the fact that $\frac{d}{dy}\cos^{-1} y = \frac{-1}{\sqrt{1-y^2}}$.

Example 4.29. Let X be a random variable with PDF

$$f_X(x) = ae^x e^{-ae^x}.$$

Let $Y = e^X$, and find $f_Y(y)$.

Solution. We first note that

$$\begin{aligned} F_Y(y) = \mathbb{P}[Y \leq y] &= \mathbb{P}[e^X \leq y] \\ &= \mathbb{P}[X \leq \log y] = \int_{-\infty}^{\log y} ae^x e^{-ae^x}\, dx. \end{aligned}$$

To find the PDF, we recall the fundamental theorem of calculus. This gives us

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}\int_{-\infty}^{\log y} ae^x e^{-ae^x}\, dx \\ &= \left(\frac{d}{dy}\log y\right)\left(\frac{d}{d\log y}\int_{-\infty}^{\log y} ae^x e^{-ae^x}\, dx\right) \\ &= \frac{1}{y} ae^{\log y} e^{-ae^{\log y}} = ae^{-ay}. \end{aligned}$$

Closing remark. The transformation of random variables is a fundamental technique in data science. The approach we have presented is the most rudimentary yet the most intuitive. The key is to visualize the transformation and how the random samples are allocated after the transformation. Note that the density of the random samples is related to the slope of the CDF. Therefore, if the transformation maps many samples to similar values, the slope of the CDF will be steep. Once you understand this picture, the transformation will be a lot easier to understand.

Is it possible to replace the paper-and-pencil derivation of a transformation with a computer? If the objective is to transform random realizations, then the answer is yes because your goal is to transform numbers to numbers, which can be done on a computer. For example, transforming a sample x_1 to $\sqrt{x_1}$ is straightforward on a computer. However, if the objective is to derive the theoretical expression of the PDF, then the answer is no. Why might we want to derive the theoretical PDF? We might want to analyze the mean, variance, or other statistical properties. We may also want to reverse-engineer and determine a transformation that can yield a specific PDF. This would require a paper-and-pencil derivation. In what follows, we will discuss a handy application of the transformations.

What are the rules of thumb for transformation of random variables?

- Always find the CDF $F_Y(y) = \mathbb{P}[g(X) \leq y]$. Ask yourself: What are the values of X such that $g(X) \leq y$? Think of the cosine example.

- Sometimes you do not need to solve for $F_Y(y)$ explicitly. The fundamental theorem of calculus can help you find $f_Y(y)$.
- Draw pictures. Ask yourself whether you need to squeeze or stretch the samples.

4.8 Generating Random Numbers

Most scientific computing software nowadays has built-in random number generators. For common types of random variables, e.g., Gaussian or exponential, these random number generators can easily generate numbers according to the chosen distribution. However, if we are given an arbitrary PDF (or PMF) that is not among the list of predefined distributions, how can we generate random numbers according to the PDF or PMF we want?

4.8.1 General principle

Generating random numbers according to the desired distribution can be formulated as an inverse problem. Suppose that we can generate uniformly random numbers according to Uniform(0,1). This is a fragile assumption, and this process can be done on almost all computers today. Let us call this random variable U and its realization u. Suppose that we also have a desired distribution $f_X(x)$ (and its CDF $F_X(x)$). We can put the two random variables U and X on the two axes of **Figure 4.35**, yielding an input-output relationship. The inverse problem is: By using what transformation g, such that $X = g(U)$, can we make sure that X is distributed according to $f_X(x)$ (or $F_X(x)$)?

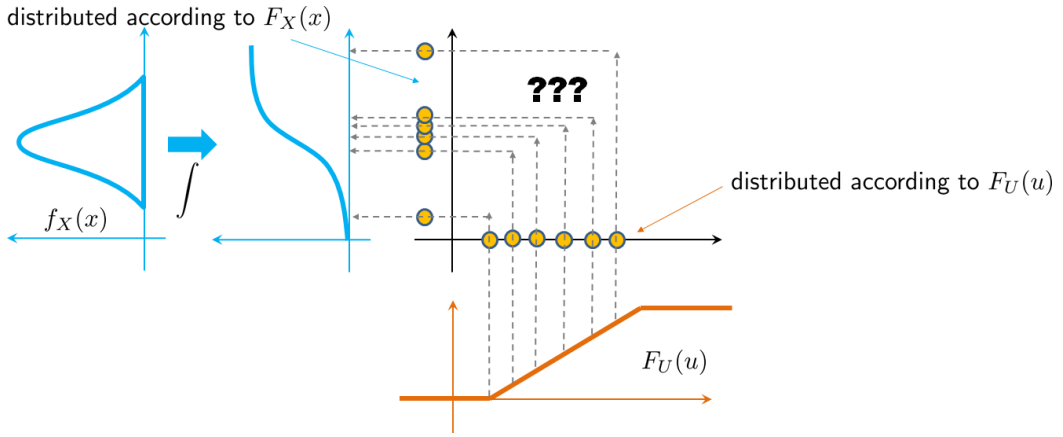


Figure 4.35: Generating random numbers according to a known CDF. The idea is to first generate a uniform(0,1) random variable, then do an inverse mapping F_X^{-1}.

Theorem 4.12. *The transformation g that can turn a uniform random variable into*

a random variable following a distribution $F_X(x)$ is given by

$$g(u) = F_X^{-1}(u). \tag{4.35}$$

That is, if $g = F_X^{-1}$, then $g(U)$ will be distributed according to f_X (or F_X).

Proof. First, we know that if $U \sim \text{Uniform}(0,1)$, then $f_U(u) = 1$ for $0 \le u \le 1$, so

$$F_U(u) = \int_{-\infty}^{u} f_U(u)\, du = u,$$

for $0 \le u \le 1$. Let $g = F_X^{-1}$ and define $Y = g(U)$. Then the CDF of Y is

$$\begin{aligned} F_Y(y) = \mathbb{P}[Y \le y] &= \mathbb{P}[g(U) \le y] \\ &= \mathbb{P}[F_X^{-1}(U) \le y] \\ &= \mathbb{P}[U \le F_X(y)] = F_X(y). \end{aligned}$$

Therefore, we have shown that the CDF of Y is the CDF of X. □

The theorem above states that if we want a distribution F_X, then the transformation should be $g = F_X^{-1}$. This suggests a two-step process for generating random numbers.

How do we generate random numbers from an arbitrary distribution F_X?

- Step 1: Generate a random number $U \sim \text{Uniform}(0,1)$.
- Step 2: Let
$$Y = F_X^{-1}(U). \tag{4.36}$$
Then the distribution of Y is F_X.

4.8.2 Examples

Example 4.30. How can we generate Gaussian random numbers with mean μ and variance σ^2 from uniform random numbers?

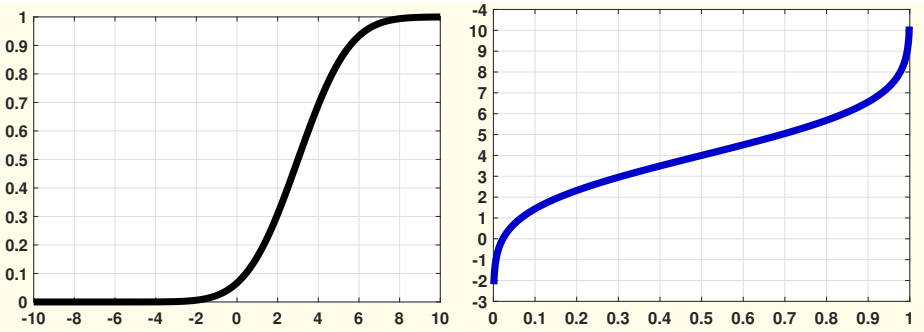
First, we generate $U \sim \text{Uniform}(0,1)$. The CDF of the ideal distribution is

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Therefore, the transformation g is

$$g(U) = F_X^{-1}(U) = \sigma\Phi^{-1}(U) + \mu.$$

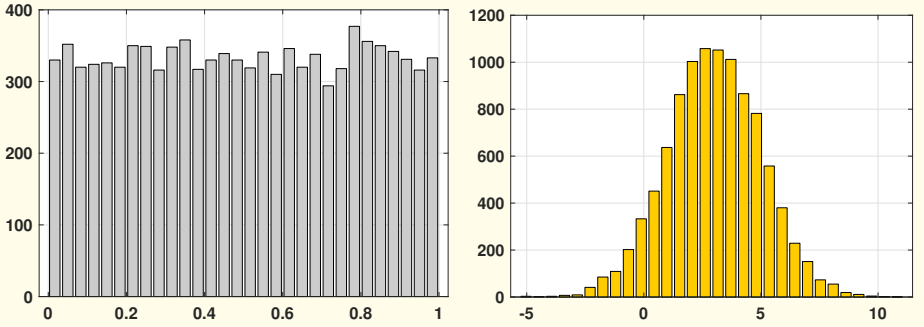
In **Figure 4.36**, we plot the CDF of F_X and the transformation g.



(a) $F_X(\cdot)$ (b) $g(\cdot)$

Figure 4.36: To generate random numbers according to Gaussian$(0, 1)$, we plot its CDF in (a) and the transformation g in (b).

To visualize the random variables before and after the transformation, we plot the histograms in **Figure 4.37**.



(a) PDF of U (b) PDF of $g(U)$

Figure 4.37: (a) PDF of the uniform random variable. (b) The PDF of the transformed random variable.

The MATLAB and Python codes used to generate the histograms above are shown below.

```
% MATLAB code to generate Gaussian from uniform
mu    = 3;
sigma = 2;
U     = rand(10000,1);
gU    = sigma*icdf('norm',U,0,1)+mu;
figure; hist(U);
figure; hist(gU);
```

```
# Python code to generate Gaussian from uniform
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
```

```
mu = 3
sigma = 2
U  = stats.uniform.rvs(0,1,size=10000)
gU = sigma*stats.norm.ppf(U)+mu
plt.hist(U); plt.show()
plt.hist(gU); plt.show()
```

Example 4.31. How can we generate exponential random numbers with parameter λ from uniform random numbers?

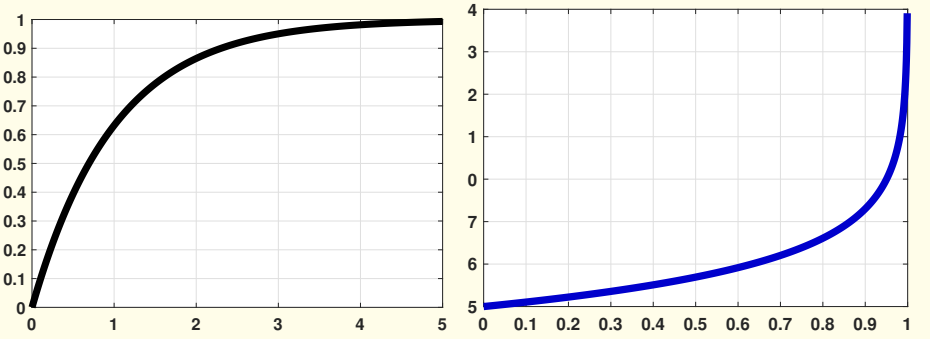
First, we generate $U \sim \text{Uniform}(0,1)$. The CDF of the ideal distribution is

$$F_X(x) = 1 - e^{-\lambda x}.$$

Therefore, the transformation g is

$$g(U) = F_X^{-1}(U) = -\frac{1}{\lambda}\log(1-U).$$

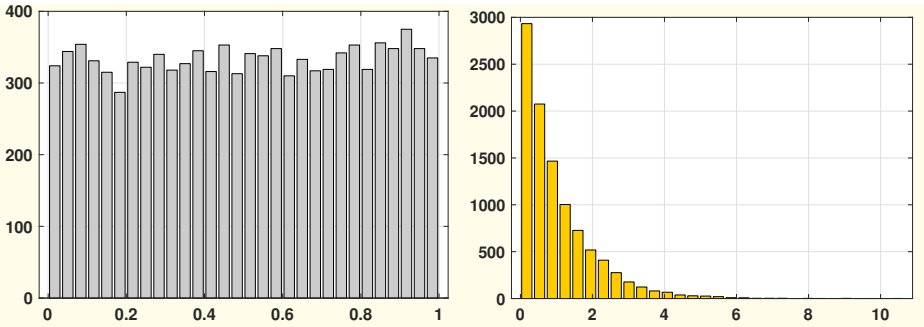
The CDF of the exponential random variable and the transformation g are shown in **Figure 4.38**.



(a) $F_X(\cdot)$ (b) $g(\cdot)$

Figure 4.38: To generate random numbers according to Exponential(1), we plot its CDF in (a) and the transformation g in (b).

The PDF of the uniform random variable U and the PDF of the transformed variable $g(U)$ are shown in **Figure 4.39**.



(a) PDF of U (b) PDF of $g(U)$

Figure 4.39: (a) PDF of the uniform random variable. (b) The PDF of the transformed random variable.

The MATLAB and Python codes for this transformation are shown below.

```
% MATLAB code to generate exponential random variables
lambda = 1;
U      = rand(10000,1);
gU     = -(1/lambda)*log(1-U);
```

```
# Python code to generate exponential random variables
import numpy as np
import scipy.stats as stats

lambd = 1;
U     = stats.uniform.rvs(0,1,size=10000)
gU    = -(1/lambd)*np.log(1-U)
```

Example 4.32. How can we generate the 4 integers $1, 2, 3, 4$, according to the histogram $[0.1\ 0.5\ 0.3\ 0.1]$, from uniform random numbers?

First, we generate $U \sim \text{Uniform}(0, 1)$. The CDF of the ideal distribution is

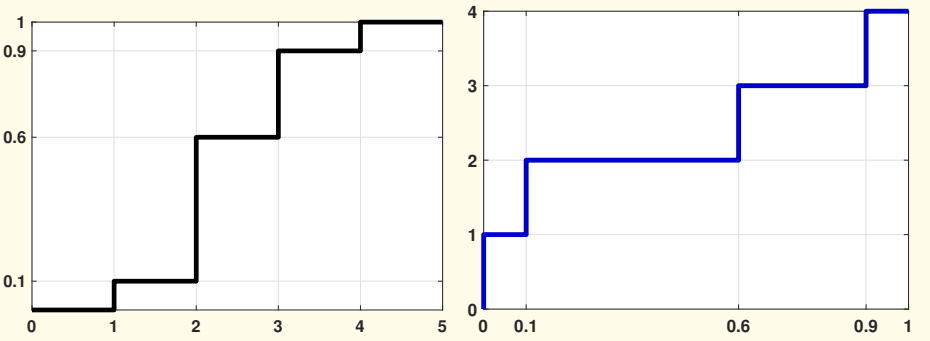
$$F_X(x) = \begin{cases} 0.1, & x = 1, \\ 0.1 + 0.5 = 0.6, & x = 2, \\ 0.1 + 0.5 + 0.3 = 0.9, & x = 3, \\ 0.1 + 0.5 + 0.3 + 0.1 = 1.0, & x = 4. \end{cases}$$

This CDF is not invertible. However, we can still define the "inverse" mapping

as

$$g(U) = F_X^{-1}(U)$$
$$= \begin{cases} 1, & 0.0 \leq U \leq 0.1, \\ 2, & 0.1 < U \leq 0.6, \\ 3, & 0.6 < U \leq 0.9, \\ 4, & 0.9 < U \leq 1.0. \end{cases}$$

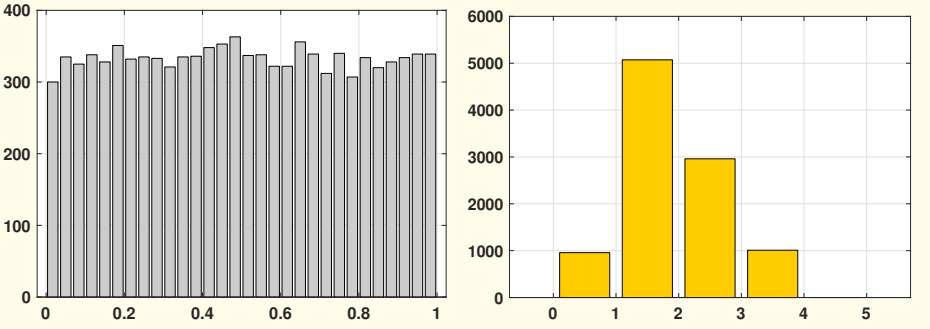
For example, if $0.1 < U \leq 0.6$, then on the black curve shown in **Figure 4.40**(a), we are looking at the second vertical line from the left. This will go to "2" on the x-axis. Therefore, the inversely mapped value is 2 for $0.1 < U \leq 0.6$.



(a) $F_X(\cdot)$ (b) $g(\cdot)$

Figure 4.40: To generate random numbers according to a predefined histogram, we first define the CDF in (a) and the corresponding transformation in (b).

The PDFs of the transformed variables, before and after, are shown in **Figure 4.41**.



(a) PDF of U (b) PDF of $g(U)$

Figure 4.41: (a) PDF of the uniform random variable. (b) The PDF of the transformed random variable.

In MATLAB, the above PDFs can be plotted using the commands below. In Python, we need to use the logical comparison `np.logical_and` to identify the indices. An alternative is to use `gU[((U<=0.5)*(U>=0.0)).astype(np.bool)]=1`.

```
% MATLAB code to generate the desired random variables
U  = rand(10000,1);
gU = zeros(10000,1);
gU((U>=0)  & (U<=0.1)) = 1;
gU((U>0.1) & (U<=0.6)) = 2;
gU((U>0.6) & (U<=0.9)) = 3;
gU((U>0.9) & (U<=1))   = 4;
```

```
# Python code to generate the desired random variables
import numpy as np
import scipy.stats as stats

U  = stats.uniform.rvs(0,1,size=10000)
gU = np.zeros(10000)
gU[np.logical_and(U >= 0.0, U <= 0.1)] = 1
gU[np.logical_and(U > 0.1, U <= 0.6)]  = 2
gU[np.logical_and(U > 0.6, U <= 0.9)]  = 3
gU[np.logical_and(U > 0.9, U <= 1)]    = 4
```

4.9 Summary

Let us summarize this chapter by revisiting the four bullet points from the beginning of the chapter.

- **Definition of a continuous random variable**. Continuous random variables are *measured* by lengths, areas, and volumes, which are all defined by integrations. This makes them different from discrete random variables, which are measured by counts (and summations). Because of the different measures being used to define random variables, we consequently have different ways of defining expectation, variance, moments, etc., all in terms of integrations.
- **Unification of discrete and continuous random variables**. The unification is done by the CDF. The CDF of a discrete random variable can be written as a train of step functions. After taking the derivative, we will obtain the PDF, which is a train of impulses.
- **Origin of Gaussian random variables**. The origin of the Gaussian random variable lies in the fact that many observable events in engineering are sums of independent events. The summation of independent random variables is equivalent to taking convolutions of the PDFs. At the limit, they will converge to a bell-shaped function, which is the Gaussian. Gaussians are everywhere because we observe sums more often than we observe individual states.
- **Transformation of random variables**. Transformation of random variables is done in the CDF space. The transformation can be used to generate random numbers

according to a predefined distribution. Specifically, if we want to generate random numbers according to F_X, then the transformation is $g = F_X^{-1}$.

4.10 Reference

PDF, CDF, expectation

4-1 Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd Edition, 2008. Chapter 3.1, 3.2.

4-2 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Chapter 4.1 - 4.3.

4-3 Athanasios Papoulis and S. Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th Edition, 2001. Chapter 4.

4-4 John A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, 2006. Chapter 4.1, 4.2, 5.1, 5.3, 5.5.

4-5 Sheldon Ross, *A First Course in Probability*, Prentice Hall, 8th Edition, 2010. Chapter 4.10, 5.1, 5.2, 5.3.

4-6 Henry Stark and John Woods, *Probability and Random Processes With Applications to Signal Processing*, Prentice Hall, 3rd edition, 2001. Chapter 2.4, 2.5, 4.1, 4.4.

Gaussian random variables

4-7 Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd Edition, 2008. Chapter 3.3.

4-8 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Chapter 4.4.

4-9 Sheldon Ross, *A First Course in Probability*, Prentice Hall, 8th Edition, 2010. Chapter 5.4.

4-10 Mark D. Ward and Ellen Gundlach, *Introduction to Probability*, W.H. Freeman and Company, 2016. Chapter 35.

Transformation of random variables

4-11 Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd Edition, 2008. Chapter 4.1.

4-12 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Chapter 4.5.

4-13 Athanasios Papoulis and S. Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th Edition, 2001. Chapter 5.

4-14 John A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, 2006. Chapter 5.4.

4-15 Sheldon Ross, *A First Course in Probability*, Prentice Hall, 8th Edition, 2010. Chapter 5.7.

4-16 Henry Stark and John Woods, *Probability and Random Processes With Applications to Signal Processing*, Prentice Hall, 3rd edition, 2001. Chapter 3.1, 3.2.

Advanced probability textbooks

4-17 William Feller, *An Introduction to Probability Theory and Its Applications*, Wiley and Sons, 3rd Edition, 1950.

4-18 Andrey Kolmogorov, *Foundations of the Theory of Probability*, 2nd English Edition, Dover 2018. (Translated from Russian to English. Originally published in 1950 by Chelsea Publishing Company New York.)

4.11 Problems

Exercise 1. (Video Solution)
Let X be a Gaussian random variable with $\mu = 5$ and $\sigma^2 = 16$.

(a) Find $\mathbb{P}[X > 4]$ and $\mathbb{P}[2 \leq X \leq 7]$.

(b) If $\mathbb{P}[X < a] = 0.8869$, find a.

(c) If $\mathbb{P}[X > b] = 0.1131$, find b.

(d) If $\mathbb{P}[13 < X \leq c] = 0.0011$, find c.

Exercise 2. (Video Solution)
Compute $\mathbb{E}[Y]$ and $\mathbb{E}[Y^2]$ for the following random variables:

(a) $Y = A\cos(\omega t + \theta)$, where $A \sim \mathcal{N}(\mu, \sigma^2)$.

(b) $Y = a\cos(\omega t + \Theta)$, where $\Theta \sim \text{Uniform}(0, 2\pi)$.

(c) $Y = a\cos(\omega T + \theta)$, where $T \sim \text{Uniform}\left(-\frac{\pi}{\omega}, \frac{\pi}{\omega}\right)$.

Exercise 3. (Video Solution)
Consider a CDF

$$F_X(x) = \begin{cases} 0, & \text{if } x < -1, \\ 0.5, & \text{if } -1 \leq x < 0, \\ (1+x)/2, & \text{if } 0 \leq x < 1, \\ 1, & \text{otherwise.} \end{cases}$$

(a) Find $\mathbb{P}[X < -1]$, $\mathbb{P}[-0.5 < X < 0.5]$ and $\mathbb{P}[X > 0.5]$.

(b) Find $f_X(x)$.

Exercise 4. (Video Solution)
A random variable X has CDF:

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1 - \frac{1}{4}e^{-2x}, & \text{if } x \geq 0. \end{cases}$$

(a) Find $\mathbb{P}[X \leq 2]$, $\mathbb{P}[X = 0]$, $\mathbb{P}[X < 0]$, $\mathbb{P}[2 < X < 6]$ and $\mathbb{P}[X > 10]$.

(b) Find $f_X(x)$.

Exercise 5. (Video Solution)
A random variable X has PDF

$$f_X(x) = \begin{cases} cx(1 - x^2), & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Find c, $F_X(x)$, and $\mathbb{E}[X]$.

Exercise 6. (Video Solution)
A continuous random variable X has a cumulative distribution

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 0.5 + c\sin^2(\pi x/2), & 0 \leq x \leq 1, \\ 1, & x > 1. \end{cases}$$

(a) What values can c assume?

(b) Find $f_X(x)$.

Exercise 7. (Video Solution)
A continuous random variable X is uniformly distributed in $[-2, 2]$.

(a) Let $Y = \sin(\pi X/8)$. Find $f_Y(y)$.

(b) Let $Z = -2X^2 + 3$. Find $f_Z(z)$.

Hint: Compute $F_Y(y)$ from $F_X(x)$, and use $\frac{d}{dy}\sin^{-1} y = \frac{1}{\sqrt{1-y^2}}$.

Exercise 8.
Let $Y = e^X$.

(a) Find the CDF and PDF of Y in terms of the CDF and PDF of X.

(b) Find the PDF of Y when X is a Gaussian random variable. In this case, Y is said to be a lognormal random variable.

Exercise 9.
The random variable X has the PDF

$$f_X(x) = \begin{cases} \frac{1}{2\sqrt{x}}, & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Let Y be a new random variable

$$Y = \begin{cases} 0, & X < 0, \\ \sqrt{X}, & 0 \le X \le 1, \\ 1, & X > 1. \end{cases}$$

Find $F_Y(y)$ and $f_Y(y)$, for $-\infty < y < \infty$.

Exercise 10.
A random variable X has the PDF

$$f_X(x) = \begin{cases} 2xe^{-x^2}, & x \ge 0, \\ 0, & x < 0. \end{cases}$$

Let

$$Y = g(X) = \begin{cases} 1 - e^{-X^2}, & X \ge 0, \\ 0, & X < 0. \end{cases}$$

Find the PDF of Y.

Exercise 11.
A random variable X has the PDF

$$f_X(x) = \frac{1}{2}e^{-|x|}, \qquad -\infty < x < \infty.$$

Let $Y = g(X) = e^{-X}$. Find the PDF of Y.

Exercise 12.
A random variable X has the PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}, \qquad -\infty < x < \infty.$$

Find the PDF of Y where

$$Y = g(X) = \begin{cases} X, & |X| > K, \\ -X, & |X| < K. \end{cases}$$

Exercise 13.
A random variable X has the PDF

$$f_X(x) = \frac{1}{x^2\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \qquad -\infty < x < \infty.$$

Let $Y = g(X) = \frac{1}{X}$. Find the PDF of Y.

Exercise 14.
A random variable X has the CDF

$$F_X(x) = \begin{cases} 0, & x < 0, \\ x^\alpha, & 0 \le x \le 1, \\ 1, & x > 1, \end{cases}$$

with $\alpha > 0$. Find the CDF of Y if

$$Y = g(X) = -\log X.$$

Exercise 15.
Energy efficiency is an important aspect of designing electrical systems. In some modern buildings (e.g., airports), traditional escalators are being replaced by a new type of "smart" escalator which can automatically switch between a normal operating mode and a standby mode depending on the flow of pedestrians.

(a) The arrival of pedestrians can be modeled as a Poisson random variable. Let N be the number of arrivals, and let λ be the arrival rate (people per minute). For a period of t minutes, show that the probability that there are n arrivals is

$$\mathbb{P}(N = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

(b) Let T be a random variable denoting the interarrival time (i.e., the time between two consecutive arrivals). Show that

$$\mathbb{P}(T > t) = e^{-\lambda t}.$$

Also, determine $F_T(t)$ and $f_T(t)$. Sketch $f_T(t)$.

(Hint: Note that $\mathbb{P}(T > t) = \mathbb{P}(\text{no arrival in } t \text{ minutes})$.)

(c) Suppose that the escalator will go into standby mode if there are no pedestrians for $t_0 = 30$ seconds. Let Y be a random variable denoting the amount of time that the escalator is in standby mode. That is, let

$$Y = \begin{cases} 0, & \text{if } T \leq t_0, \\ T - t_0, & \text{if } T > t_0. \end{cases}$$

Find $\mathbb{E}[Y]$.

Chapter 5

Joint Distributions

When you go to a concert hall, sometimes you may want to see a solo violin concert, but other times you may want to see a symphony. Symphonies are appealing because many instruments are playing together. Random variables are similar. While single random variables are useful for modeling simple events, we use multiple random variables to describe complex events. The multiple random variables can be either independent or correlated. When many random variables are present in the problem, we enter the subject of **joint distribution**.

What are joint distributions?

In the simplest sense, joint distributions are extensions of the PDFs and PMFs we studied in the previous chapters. We summarize them as follows.

> Joint distributions are **high-dimensional** PDFs (or PMFs or CDFs).

What do we mean by a high-dimensional PDF? We know that a single random variable is characterized by a 1-dimensional PDF $f_X(x)$. If we have a pair of random variables, then we use a 2-dimensional function $f_{X,Y}(x,y)$, and if we have a triplet of random variables, we use a 3-dimensional function $f_{X,Y,Z}(x,y,z)$. In general, the dimensionality of the PDF grows as the number of variables:

$$\underbrace{f_X(x)}_{\text{one variable}} \implies \underbrace{f_{X_1,X_2}(x_1,x_2)}_{\text{two variables}} \implies \cdots \implies \underbrace{f_{X_1,\ldots,X_N}(x_1,\ldots,x_N)}_{N \text{ variables}}.$$

For busy engineers like us, $f_{X_1,\ldots,X_N}(x_1,\ldots,x_N)$ is not a friendly notation. A more concise way to write $f_{X_1,\ldots,X_N}(x_1,\ldots,x_N)$ is to define a *vector* of random variables $\boldsymbol{X} = [X_1, X_2, \ldots, X_N]^T$ with a vector of states $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^T$, and to define the PDF as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1,\ldots,X_N}(x_1,\ldots,x_N).$$

Under what circumstance will we encounter creatures like $f_{\boldsymbol{X}}(\boldsymbol{x})$? Believe it or not, these high-dimensional PDFs are *everywhere*. In 2010, computer-vision scientists created the ImageNet dataset, containing 14 million images with ground-truth class labels. This enormous dataset has enabled a great blossoming of machine learning over the past several

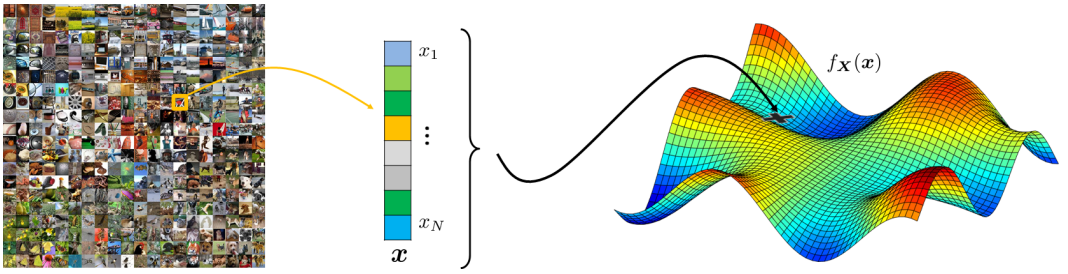


Figure 5.1: Joint distributions are ubiquitous in modern data analysis. For example, an image from a dataset can be represented by a high-dimensional vector $\boldsymbol{x}$. Each vector has a certain probability of being present. This probability is described by the high-dimensional joint PDF $f_{\boldsymbol{X}}(\boldsymbol{x})$. The goal of this chapter is to understand the properties of this $f_{\boldsymbol{X}}$.

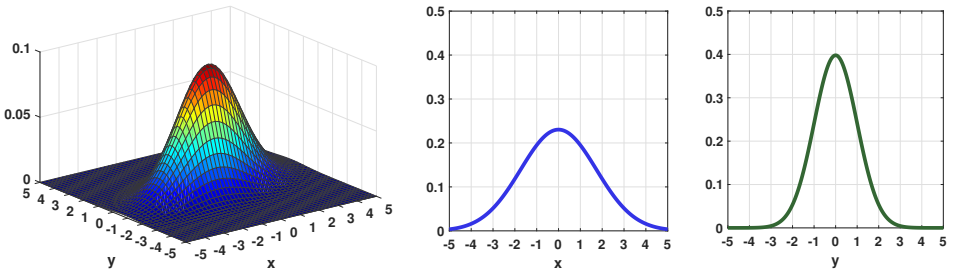


Figure 5.2: A 2-dimensional PDF $f_{X,Y}(x, y)$ of a pair of random variables (X, Y) and their respective 1D PDFs $f_X(x)$ and $f_Y(y)$.

decades, in which many advances in deep learning have been made. Fundamentally, the ImageNet dataset provides a large collection of *samples* drawn from a latent distribution that is high-dimensional. Each sample in the ImageNet dataset is a $224 \times 224 \times 3$ image (the three numbers stand for the image's height, width, and color). If we convert this image into a vector, then the sample will have a dimension of $224 \times 224 \times 3 = 150{,}528$. In other words, the sample is a vector $\boldsymbol{x} \in \mathbb{R}^{150528 \times 1}$. The probability of obtaining a particular sample $\boldsymbol{x}$ is determined by probability density function $f_{\boldsymbol{X}}(\boldsymbol{x})$. For example, it is more likely to get an image containing trees than one containing a Ferrari. The manifold generated by $f_{\boldsymbol{X}}(\boldsymbol{x})$ can be extremely complex, as illustrated in **Figure 5.1**.

The story of ImageNet is just one of the many instances for which we use a joint distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$. Joint distributions are ubiquitous. If you do data science, you *must* understand joint distributions. However, extending a 1-dimensional function $f_X(x)$ to a 2-dimensional function $f_{X,Y}(x, y)$ and then to a N-dimensional function $f_{\boldsymbol{X}}(\boldsymbol{x})$ is not trivial. The goal of this chapter is to guide you through these important steps.

Plan of Part 1 of this chapter: Two variables

This chapter is broadly divided into two halves. In the first half, we will look at **a pair of random variables**.

- **Definition of** $f_{X,Y}(x, y)$. The first thing we need to learn is the definition of a joint distribution with two variables. Since we have two variables, the **joint probability density function** (or probability mass function) is a 2-dimensional function. A point

on this 2D function is the probability density evaluated by a pair of variables $X = x$ and $Y = y$, as illustrated in **Figure 5.2**. However, how do we formally define this 2D function? How is it related to the probability measure? Is there a way we can retrieve $f_X(x)$ and $f_Y(y)$ from $f_{X,Y}(x, y)$, as illustrated on the right-hand sides of **Figure 5.2**? These questions will be answered in Section 5.1.

- **Joint expectation** $\mathbb{E}[XY]$. When we have a pair of random variables, how should we define the expectation? In Section 5.2, we will show that the most natural way to define the joint expectation is in terms of $\mathbb{E}[XY]$, i.e., the expectation of the product. There is a surprising and beautiful connection between this "expectation of the product" and the cosine angle between two vectors, thereby showing that $\mathbb{E}[XY]$ is the **correlation** between X and Y.
- The reason for studying a pair of random variables is to spell out the cause-effect relationship between the variables. This cannot be done without conditional distributions; this will be explained in Section 5.3. Conditional distributions provide an extremely important computational tool for decoupling complex events into simpler events. Such decomposition allows us to solve difficult joint expectation problems via simple **conditional expectations**; this subject will be covered in Section 5.4.
- If you recall our discussions about the origin of a Gaussian random variable, we claimed that the PDF of $X + Y$ is the **convolution** between f_X and f_Y. Why is this so? We will answer this question in terms of joint distributions in Section 5.5.

Plan of Part 2 of this chapter: N variables

The second half of the chapter focuses on the general case of N random variables. This requires the definitions of a random vector $\boldsymbol{X} = [X_1, \ldots, X_N]^T$, a joint distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$, and the corresponding expectations $\mathbb{E}[\boldsymbol{X}]$. To make our discussions concrete, we will focus on the case of **high-dimensional Gaussian** random variables and discuss the following topics.

- **Covariance matrices/correlation matrices**. If a pair of random variables can define the correlation through the expectation of the product $\mathbb{E}[X_1X_2]$, then for a vector of random variables we can consider a matrix of correlations in the form

$$\boldsymbol{R} = \begin{bmatrix} \mathbb{E}[X_1X_1] & \mathbb{E}[X_1X_2] & \cdots & \mathbb{E}[X_1X_N] \\ \mathbb{E}[X_2X_1] & \mathbb{E}[X_2X_2] & \cdots & \mathbb{E}[X_2X_N] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_NX_1] & \mathbb{E}[X_NX_2] & \cdots & \mathbb{E}[X_NX_N] \end{bmatrix}.$$

  What are the properties of the matrix? How does it affect the shape of the high-dimensional Gaussian? If we have a dataset of vectors, how do we estimate this matrix from the data? We will answer these questions in Section 5.6 and Section 5.7.
- **Principal-component analysis**. Given the covariance matrix, we can perform some very useful data analyses, such as the principal-component analysis in Section 5.8. The question we will ask is: Among the many components, which one is the principal component? If we can find the principal component(s), we can effectively perform dimensionality reduction by projecting a high-dimensional vector into low-dimensional representations. We will introduce an application for face detection.

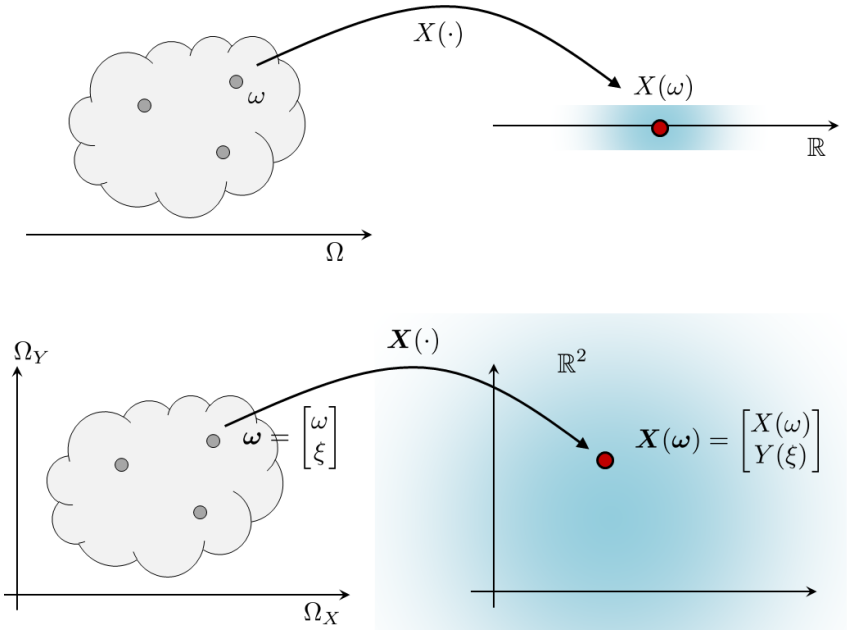


Figure 5.3: When there is a pair of random variables, we can regard the sample space as a set of coordinates. The random variables are 2D mappings from a coordinate $\boldsymbol{\omega}$ in $\Omega_X \times \Omega_Y$ to another coordinate $\boldsymbol{X}(\boldsymbol{\omega})$ in $\mathbb{R}^2$.

5.1 Joint PMF and Joint PDF

Probability is a measure of the size of a set. This principle applies to discrete random variables, continuous random variables, single random variables, and multiple random variables. In situations with a pair of random variables, the measure should be applied to the coordinate (X, Y) represented by the random variables X and Y. Consequently, when measuring the probability, we either count these coordinates or integrate the area covered by these coordinates. In this section, we formalize this notion of measuring 2D events.

5.1.1 Probability measure in 2D

Consider two random variables X and Y. Let the sample space of X and Y be Ω_X and Ω_Y, respectively. Define the **Cartesian product** of Ω_X and Ω_Y as $\Omega_X \times \Omega_Y = \{(x, y) \mid x \in \Omega_X \text{ and } y \in \Omega_Y\}$. That is, $\Omega_X \times \Omega_Y$ contains all possible pairs (X, Y).

Example 5.1. If $\Omega_X = \{1, 2\}$ and $\Omega_Y = \{4, 5\}$, then $\Omega_X \times \Omega_Y = \{(1, 4), (1, 5), (2, 4), (2, 5)\}$.

Example 5.2. If $\Omega_X = [3,4]$ and $\Omega_Y = [1,2]$, then $\Omega_X \times \Omega_Y =$ a rectangle with two diagonal vertices as $(3,1)$ and $(4,2)$.

Random variables are mappings from the sample space to the real line. If $\omega \in \Omega_X$ is mapped to $X(\omega) \in \mathbb{R}$, and $\xi \in \Omega_Y$ is mapped to $Y(\xi) \in \mathbb{R}$, then a coordinate $\boldsymbol{\omega} = (\omega, \xi)$ in the sample space $\Omega_X \times \Omega_Y$ should be mapped to a coordinate $(X(\omega), Y(\xi))$ in the 2D plane.

$$\boldsymbol{\omega} \overset{\text{def}}{=} \begin{bmatrix} \omega \\ \xi \end{bmatrix} \longmapsto \begin{bmatrix} X(\omega) \\ Y(\xi) \end{bmatrix} \overset{\text{def}}{=} \boldsymbol{X}(\boldsymbol{\omega}).$$

We denote such a vector-to-vector mapping as $\boldsymbol{X}(\cdot) : \Omega_X \times \Omega_Y \to \mathbb{R} \times \mathbb{R}$, as illustrated in **Figure 5.3**.

Therefore, if we have an event $\mathcal{A} \in \mathbb{R}^2$, the probability that $\mathcal{A}$ happens is

$$\begin{aligned} \mathbb{P}[\mathcal{A}] &= \mathbb{P}[\{\boldsymbol{\omega} \mid \boldsymbol{X}(\boldsymbol{\omega}) \in \mathcal{A}\}] \\ &= \mathbb{P}\left[\left\{ \begin{bmatrix} \omega \\ \xi \end{bmatrix} \;\middle|\; \begin{bmatrix} X(\omega) \\ Y(\xi) \end{bmatrix} \in \mathcal{A} \right\}\right] \\ &= \mathbb{P}\left[\left\{ \begin{bmatrix} \omega \\ \xi \end{bmatrix} \in \boldsymbol{X}^{-1}(\mathcal{A}) \right\}\right] \\ &= \mathbb{P}[\boldsymbol{\omega} \in \boldsymbol{X}^{-1}(\mathcal{A})]. \end{aligned}$$

In other words, we take the coordinate $\boldsymbol{X}(\boldsymbol{\omega})$ and find its inverse image $\boldsymbol{X}^{-1}(\mathcal{A})$. The size of this inverse image $\boldsymbol{X}^{-1}(\mathcal{A})$ in the sample space $\Omega_X \times \Omega_Y$ is then the probability. We summarize this general principle as follows.

How to measure probability in 2D

For a pair of random variables $\boldsymbol{X} = (X, Y)$, the probability of an event $\mathcal{A}$ is measured in the product space $\Omega_X \times \Omega_Y$ with the size

$$\mathbb{P}[\{\boldsymbol{\omega} \mid \boldsymbol{X}^{-1}(\mathcal{A})\}].$$

This definition is quite abstract. To make it more concrete, we will look at discrete and continuous random variables.

5.1.2 Discrete random variables

Suppose that the random variables X and Y are discrete. Let $\mathcal{A} = \{X(\omega) = x,\ Y(\xi) = y\}$ be a discrete event. Then the above definition tells us that the probability of $\mathcal{A}$ is

$$\mathbb{P}[\mathcal{A}] = \mathbb{P}\left[(\omega, \xi) \;\middle|\; X(\omega) = x, \text{ and } Y(\xi) = y\right] = \underbrace{\mathbb{P}[X = x \text{ and } Y = y]}_{\overset{\text{def}}{=} p_{X,Y}(x,y)}.$$

We define this probability as the **joint probability mass function** (joint PMF) $p_{X,Y}(x,y)$.

Definition 5.1. *Let X and Y be two discrete random variables. The* **joint PMF** *of X and Y is defined as*

$$p_{X,Y}(x,y) = \mathbb{P}[X = x \text{ and } Y = y] = \mathbb{P}\Big[(\omega,\xi) \;\Big|\; X(\omega) = x, \text{ and } Y(\xi) = y\Big]. \tag{5.1}$$

We sometimes write the joint PMF as $p_{X,Y}(x,y) = \mathbb{P}[X = x,\ Y = y]$.

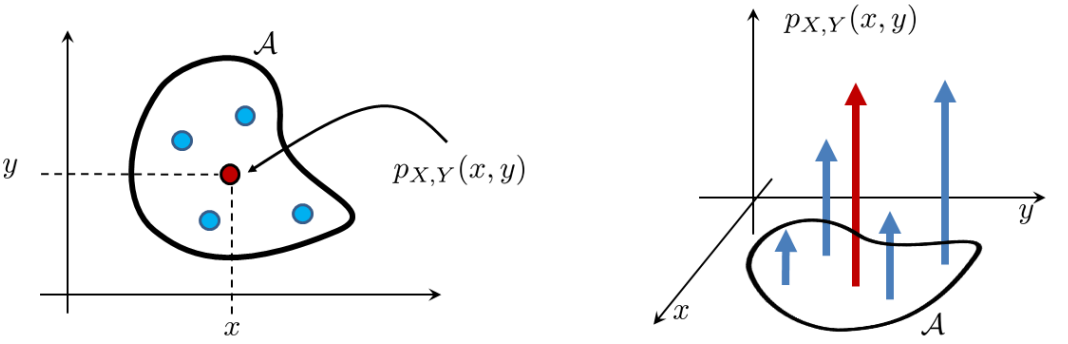


Figure 5.4: A joint PMF for a pair of discrete random variables consists of an array of impulses. To measure the size of the event $\mathcal{A}$, we sum all the impulses inside $\mathcal{A}$.

Figure 5.4 shows a graphical portrayal of the joint PMF. In a nutshell, $p_{X,Y}(x,y)$ can be considered as a 2D extension of a single variable PMF. The probabilities are still represented by the impulses, but the domain of these impulses is now a 2D plane. If we have an event $\mathcal{A}$, then the size of the event is

$$\mathbb{P}[\mathcal{A}] = \sum_{(x,y)\in\mathcal{A}} p_{X,Y}(x,y).$$

Example 5.3. Let X be a coin flip, Y be a die. The sample space of X is $\{0,1\}$, whereas the sample space of Y is $\{1,2,3,4,5,6\}$. The joint PMF, according to our definition, is the probability $\mathbb{P}[X = x \text{ and } Y = y]$, where x takes a binary state and Y takes one of the 6 states. The following table summarizes all the 12 states of the joint distribution.

	Y					
	1	2	3	4	5	6
X = 0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
X = 1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

In this table, since there are 12 coordinates, and each coordinate has an equal chance of appearing, the probability for each coordinate becomes $1/12$. Therefore, the joint PMF of X and Y is

$$p_{X,Y}(x,y) = \frac{1}{12}, \quad x = 0,1, \quad y = 1,2,3,4,5,6.$$

In this example, we observe that if X and Y are not interacting with each other (formally, **independent**), the joint PMF is the product of the two individual probabilities.

Example 5.4. In the previous example, if we define $\mathcal{A} = \{X+Y = 3\}$, the probability $\mathbb{P}[\mathcal{A}]$ is

$$\begin{aligned}\mathbb{P}[\mathcal{A}] &= \sum_{(x,y)\in\mathcal{A}} p_{X,Y}(x,y) = p_{X,Y}(0,3) + p_{X,Y}(1,2) \\ &= \frac{2}{12}.\end{aligned}$$

If $\mathcal{B} = \{\min(X,Y) = 1\}$, the probability $\mathbb{P}[\mathcal{B}]$ is

$$\begin{aligned}\mathbb{P}[\mathcal{B}] &= \sum_{(x,y)\in\mathcal{B}} p_{X,Y}(x,y) \\ &= p_{X,Y}(1,1) + p_{X,Y}(1,2) + p_{X,Y}(1,3) \\ &\qquad + p_{X,Y}(1,4) + p_{X,Y}(1,5) + p_{X,Y}(1,6) \\ &= \frac{6}{12}.\end{aligned}$$

5.1.3 Continuous random variables

The continuous version of the joint PMF is called the **joint probability density function** (**joint PDF**), denoted by $f_{X,Y}(x,y)$. A joint PDF is analogous to a joint PMF. For example, integrating it will give us the probability.

Definition 5.2. *Let X and Y be two continuous random variables. The* **joint PDF** *of X and Y is a function $f_{X,Y}(x,y)$ that can be integrated to yield a probability*

$$\mathbb{P}[\mathcal{A}] = \int_{\mathcal{A}} f_{X,Y}(x,y)\, dx\, dy, \tag{5.2}$$

for any event $\mathcal{A} \subseteq \Omega_X \times \Omega_Y$.

Pictorially, we can view $f_{X,Y}$ as a 2D function where the height at a coordinate (x,y) is $f_{X,Y}(x,y)$, as can be seen from **Figure 5.5**. To compute the probability that $(X,Y) \in \mathcal{A}$, we integrate the function $f_{X,Y}$ with respect to the area covered by the set $\mathcal{A}$. For example, if the set $\mathcal{A}$ is a rectangular box $\mathcal{A} = [a,b] \times [c,d]$, then the integration becomes

$$\begin{aligned}\mathbb{P}[\mathcal{A}] &= \mathbb{P}[a \le X \le b,\ \ c \le Y \le d] \\ &= \int_c^d \int_a^b f_{X,Y}(x,y)\, dx\, dy.\end{aligned}$$

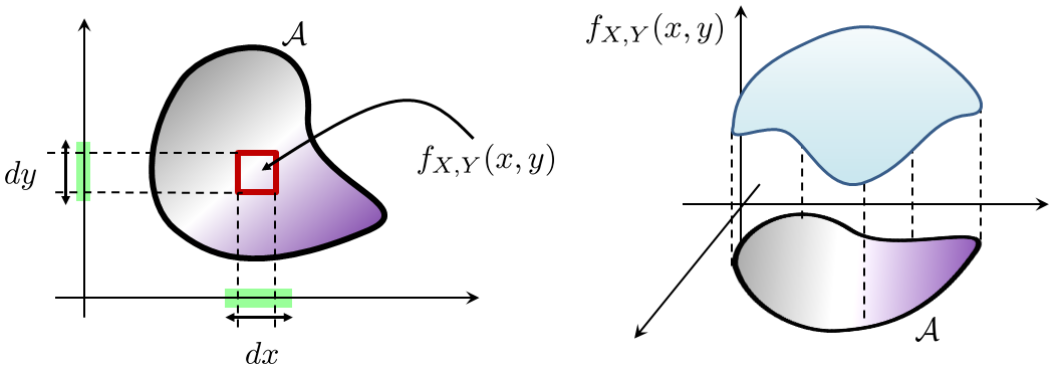


Figure 5.5: A joint PDF for a pair of continuous random variables is a surface in the 2D plane. To measure the size of the event $\mathcal{A}$, we integrate $f_{X,Y}(x,y)$ inside $\mathcal{A}$.

Example 5.5. Consider a uniform joint PDF $f_{X,Y}(x,y)$ defined on $[0,2]^2$ with $f_{X,Y}(x,y) = \frac{1}{4}$. Let $\mathcal{A} = [a,b] \times [c,d]$. Find $\mathbb{P}[\mathcal{A}]$.

Solution.

$$\begin{aligned}\mathbb{P}[\mathcal{A}] &= \mathbb{P}[a \le X \le b, \;\; c \le X \le d] \\ &= \int_c^d \int_a^b f_{X,Y}(x,y)\, dx\, dy = \int_c^d \int_a^b \frac{1}{4}\, dx\, dy = \frac{(d-c)(b-a)}{4}.\end{aligned}$$

Practice Exercise 5.1. In the previous example, let $\mathcal{B} = \{X + Y \le 2\}$. Find $\mathbb{P}[\mathcal{B}]$.

Solution.

$$\begin{aligned}\mathbb{P}[\mathcal{B}] = \int_{\mathcal{B}} f_{X,Y}(x,y)\, dx\, dy &= \int_0^2 \int_0^{2-y} f_{X,Y}(x,y)\, dx\, dy \\ &= \int_0^2 \int_0^{2-y} \frac{1}{4}\, dx\, dy = \int_0^2 \frac{2-y}{4}\, dy = \frac{1}{2}.\end{aligned}$$

Here, the limits of the integration can be determined from **Figure 5.6**. The inner integration (with respect to x) should start from 0 and end at $2-y$, which is the line defining the set $x + y \le 2$. Since the inner integration is performed for every y, we need to enumerate all the possible y's to complete the outer integration. This leads to the outer limit from 0 to 2.

5.1.4 Normalization

The normalization property of a two-dimensional PMF and PDF is the property that, when we enumerate all outcomes of the sample space, we obtain 1.

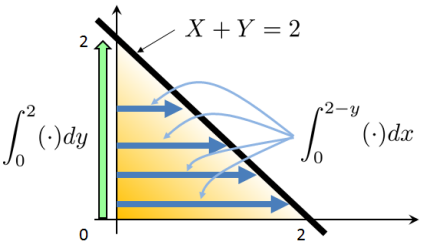


Figure 5.6: To calculate $\mathbb{P}[X+Y\le 2]$, we perform a 2D integration over a triangle.

Theorem 5.1. *Let $\Omega=\Omega_X\times\Omega_Y$. All joint PMFs and joint PDFs satisfy*

$$\sum_{(x,y)\in\Omega} p_{X,Y}(x,y)=1 \quad \text{or} \quad \int_\Omega f_{X,Y}(x,y)\,dx\,dy=1. \tag{5.3}$$

Example 5.6. Consider a joint uniform PDF defined in the shaded area $[0,3]\times[0,3]$ with PDF defined below. Find the constant c.

$$f_{X,Y}(x,y)=\begin{cases} c & \text{if } (x,y)\in[0,3]\times[0,3],\\ 0 & \text{otherwise.}\end{cases}$$

Solution. To find the constant c, we note that

$$1=\int_0^3\int_0^3 f_{X,Y}(x,y)\,dx\,dy=\int_0^3\int_0^3 c\,dx\,dy=9c.$$

Equating the two sides gives us $c=\frac{1}{9}$.

Practice Exercise 5.2. Consider a joint PDF

$$f_{X,Y}(x,y)=\begin{cases} ce^{-x}e^{-y} & 0\le y\le x<\infty,\\ 0 & \text{otherwise.}\end{cases}$$

Find the constant c. Tip: Consider the area of integration as shown in **Figure 5.7**.

Solution. There are two ways to take the integration shown in **Figure 5.7**. We choose the inner integration w.r.t. y first.

$$\int_\Omega f_{X,Y}(x,y)\,dx\,dy=\int_0^\infty\int_0^x ce^{-x}e^{-y}\,dy\,dx=\int_0^\infty ce^{-x}(1-e^{-x})=\frac{c}{2}.$$

Therefore, $c=2$.

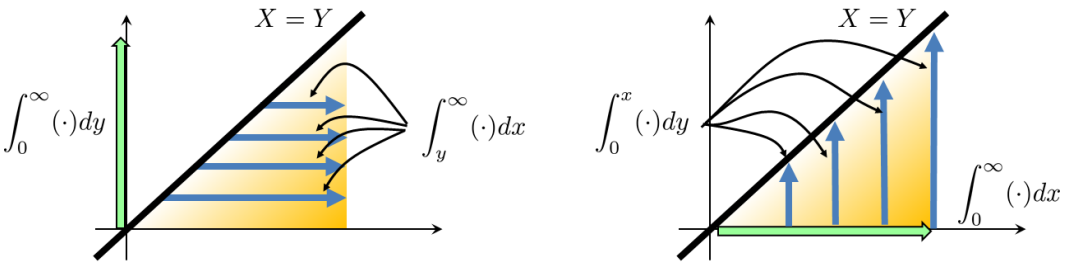


Figure 5.7: To integrate the probability $\mathbb{P}[0 \le Y \le X]$, we perform a 2D integration over a triangle. The two subfigures show the two ways of integrating the triangle. [Left] $\int\ dx$ first, and then $\int\ dy$. [Right] $\int\ dy$ first, and then $\int\ dx$.

5.1.5 Marginal PMF and marginal PDF

If we only sum / integrate for one random variable, we obtain the PMF / PDF of the other random variable. The resulting PMF / PDF is called the marginal PMF / PDF.

Definition 5.3. *The* **marginal PMF** *is defined as*

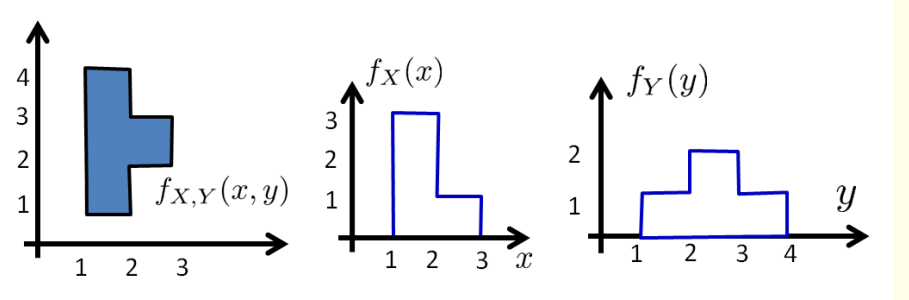
$$p_X(x) = \sum_{y \in \Omega_Y} p_{X,Y}(x,y) \quad \text{and} \quad p_Y(y) = \sum_{x \in \Omega_X} p_{X,Y}(x,y), \tag{5.4}$$

and the **marginal PDF** *is defined as*

$$f_X(x) = \int_{\Omega_Y} f_{X,Y}(x,y)\, dy \quad \text{and} \quad f_Y(y) = \int_{\Omega_X} f_{X,Y}(x,y)\, dx. \tag{5.5}$$

Since $f_{X,Y}(x,y)$ is a two-dimensional function, when integrating over y from $-\infty$ to ∞, we project $f_{X,Y}(x,y)$ onto the x-axis. Therefore, the resulting function depends on x only.

Example 5.7. Consider the joint PDF $f_{X,Y}(x,y) = \frac{1}{4}$ shown below. Find the marginal PDFs.



Solution. If we integrate over x and y, we have

$$f_X(x) = \begin{cases} 3, & \text{if } 1 < x \leq 2, \\ 1, & \text{if } 2 < x \leq 3, \\ 0, & \text{otherwise.} \end{cases} \quad \text{and} \quad f_Y(y) = \begin{cases} 1, & \text{if } 1 < x \leq 2, \\ 2, & \text{if } 2 < x \leq 3, \\ 1, & \text{if } 3 < x \leq 4, \\ 0, & \text{otherwise.} \end{cases}$$

So the marginal PDFs are the projection of the joint PDFs onto the x- and y-axes.

Practice Exercise 5.3. A joint Gaussian random variable (X, Y) has a joint PDF given by

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{((x-\mu_X)^2 + (y-\mu_Y)^2)}{2\sigma^2}\right\}.$$

Find the marginal PDFs $f_X(x)$ and $f_Y(y)$.

Solution.

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{((x-\mu_X)^2 + (y-\mu_Y)^2)}{2\sigma^2}\right\} dy \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_X)^2}{2\sigma^2}\right\} \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu_Y)^2}{2\sigma^2}\right\} dy. \end{aligned}$$

Recognizing that the last integral is equal to unity because it integrates a Gaussian PDF over the real line, it follows that

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_X)^2}{2\sigma^2}\right\}.$$

Similarly, we have

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu_Y)^2}{2\sigma^2}\right\}.$$

5.1.6 Independent random variables

Two random variables are said to be independent if and only if the joint PMF or PDF can be factorized as a product of the marginal PMF / PDFs.

Definition 5.4. *Random variables X and Y are* **independent** *if and only if*

$$p_{X,Y}(x, y) = p_X(x)\, p_Y(y), \quad \text{or} \quad f_{X,Y}(x, y) = f_X(x)\, f_Y(y).$$

This definition is consistent with the definition of independence of two events. Recall that two events A and B are independent if and only if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. Letting $A = \{X = x\}$

and $B = \{Y = y\}$, we see that if A and B are independent then $\mathbb{P}[X = x \cap Y = y]$ is the product $\mathbb{P}[X = x]\mathbb{P}[Y = y]$. This is precisely the relationship $p_{X,Y}(x, y) = p_X(x)\, p_Y(y)$.

Example 5.8. Consider two random variables with a joint PDF given by

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{(x - \mu_X)^2 + (y - \mu_Y)^2}{2\sigma^2}\right\}.$$

Are X and Y independent?

Solution. We know that

$$f_{X,Y}(x, y) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu_X)^2}{2\sigma^2}\right\}}_{f_X(x)} \times \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y - \mu_Y)^2}{2\sigma^2}\right\}}_{f_Y(y)}.$$

Therefore, the random variables X and Y are independent.

Practice Exercise 5.4. Let X be a coin and Y be a die. Then the joint PMF is given by the table below.

	Y					
	1	2	3	4	5	6
X = 0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
X = 1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

Are X and Y independent?

Solution. For any x and y, we have that

$$p_{X,Y}(x, y) = \frac{1}{12} = \underbrace{\frac{1}{2}}_{p_X(x)} \times \underbrace{\frac{1}{6}}_{p_Y(y)}.$$

Therefore, the random variables X and Y are independent.

Example 5.9. Consider two random variables X and Y with a joint PDF given by[a]

$$\begin{aligned} f_{X,Y}(x, y) \propto \exp\left\{-(x - y)^2\right\} &= \exp\left\{-x^2 + 2xy - y^2\right\} \\ &= \underbrace{\exp\left\{-x^2\right\}}_{f_X(x)} \underbrace{\exp\left\{2xy\right\}}_{\text{extra term}} \underbrace{\exp\left\{-y^2\right\}}_{f_Y(y)}. \end{aligned}$$

This PDF cannot be factorized into a product of two marginal PDFs. Therefore, the random variables are dependent.

[a] We use the notation "$\propto$" to denote "proportional to". It implies that the normalization constant is omitted.

We can extrapolate the definition of independence to multiple random variables. If there are many random variables $X_1, X_2, \ldots, X_N$, they will have a joint PDF

$$f_{X_1,\ldots,X_N}(x_1,\ldots,x_N).$$

If these random variables $X_1, X_2, \ldots, X_N$ are independent, then the joint PDF can be factorized as

$$\begin{aligned} f_{X_1,\ldots,X_N}(x_1,\ldots,x_N) &= f_{X_1}(x_1)\cdot f_{X_2}(x_2)\cdots f_{X_N}(x_N) \\ &= \prod_{n=1}^{N} f_{X_n}(x_n). \end{aligned}$$

This gives us the definition of independence for N random variables.

Definition 5.5. *A sequence of random variables $X_1, \ldots, X_N$ is* **independent** *if and only if their joint PDF (or joint PMF) can be factorized.*

$$f_{X_1,\ldots,X_N}(x_1,\ldots,x_N) = \prod_{n=1}^{N} f_{X_n}(x_n). \tag{5.6}$$

Example 5.10. Throw a die 4 times. Let X_1, X_2, X_3 and X_4 be the outcomes. Then, since these four throws are independent, the probability mass function of any quadrable (x_1, x_2, x_3, x_4) is

$$p_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4) = p_{X_1}(x_1)\, p_{X_2}(x_2)\, p_{X_3}(x_3)\, p_{X_4}(x_4).$$

For example, the probability of getting $(1, 5, 2, 6)$ is

$$p_{X_1,X_2,X_3,X_4}(1,5,2,6) = p_{X_1}(1)\, p_{X_2}(5)\, p_{X_3}(2)\, p_{X_4}(6) = \left(\frac{1}{6}\right)^4.$$

The example above demonstrates an interesting phenomenon. If the N random variables are independent, and if they all have the same distribution, then the joint PDF/PMF is just one of the individual PDFs taken to the power N. Random variables satisfying this property are known as **independent and identically distributed** random variables.

Definition 5.6 (**Independent and Identically Distributed (i.i.d.)**). *A collection of random variables $X_1, \ldots, X_N$ is called independent and identically distributed (i.i.d.) if*

- *All $X_1, \ldots, X_N$ are independent; and*
- *All $X_1, \ldots, X_N$ have the same distribution, i.e., $f_{X_1}(x) = \cdots = f_{X_N}(x)$.*

If $X_1, \ldots, X_N$ are i.i.d., we have that

$$f_{X_1,\ldots,X_N}(x_1,\ldots,x_1) = \prod_{n=1}^{N} f_{X_1}(x_n),$$

where the particular choice of X_1 is unimportant because $f_{X_1}(x) = \cdots = f_{X_N}(x)$.

Why is i.i.d. so important?

- If a set of random variables are i.i.d., then the joint PDF can be written as a product of PDFs.
- Integrating a joint PDF is difficult. Integrating a product of PDFs is much easier.

Example 5.11. Let $X_1, X_2, \ldots, X_N$ be a sequence of i.i.d. Gaussian random variables where each X_i has a PDF

$$f_{X_i}(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}.$$

The joint PDF of $X_1, X_2, \ldots, X_N$ is

$$\begin{aligned} f_{X_1,\ldots,X_N}(x_1, \ldots, x_N) &= \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x_i^2}{2}\right\} \right\} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^N \exp\left\{-\sum_{i=1}^{N} \frac{x_i^2}{2}\right\}, \end{aligned}$$

which is a function depending not on the individual values of $x_1, x_2, \ldots, x_N$ but on the sum $\sum_{i=1}^{N} x_i^2$. So we have "compressed" an N-dimensional function into a 1D function.

Example 5.12. Let θ be a deterministic number that was sent through a noisy channel. We model the noise as an additive Gaussian random variable with mean 0 and variance σ^2. Supposing we have observed measurements $X_i = \theta + W_i$, for $i = 1, \ldots, N$, where $W_i \sim \text{Gaussian}(0, \sigma^2)$, then the PDF of each X_i is

$$f_{X_i}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}.$$

Thus the joint PDF of $(X_1, X_2, \ldots, X_N)$ is

$$\begin{aligned} f_{X_1,\ldots,X_N}(x_1, \ldots, x_N) &= \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma^2}\right\} \right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{i=1}^{N} \frac{(x_i-\theta)^2}{2\sigma^2}\right\}. \end{aligned}$$

Essentially, this joint PDF tells us the probability density of seeing sample data $x_1, \ldots, x_N$.

5.1.7 Joint CDF

We now introduce the cumulative distribution function (CDF) for multiple variables.

Definition 5.7. *Let X and Y be two random variables. The* **joint CDF** *of X and Y is the function $F_{X,Y}(x,y)$ such that*

$$F_{X,Y}(x,y) = \mathbb{P}[X \le x \cap Y \le y]. \tag{5.7}$$

This definition can be more explicitly written as follows.

Definition 5.8. *If X and Y are discrete, then*

$$F_{X,Y}(x,y) = \sum_{y' \le y} \sum_{x' \le x} p_{X,Y}(x',y'). \tag{5.8}$$

If X and Y are continuous, then

$$F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(x',y')\, dx'\, dy'. \tag{5.9}$$

If the two random variables are **independent**, then we have

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} f_X(x')\, dx' \int_{-\infty}^{y} f_Y(y')\, dy' = F_X(x) F_Y(y).$$

Example 5.13. Let X and Y be two independent uniform random variables Uniform$(0,1)$. Find the joint CDF.

Solution.

$$\begin{aligned} F_{X,Y}(x,y) = F_X(x)F_Y(y) &= \int_0^x f_X(x')\, dx' \int_0^y f_Y(y')\, dy' \\ &= \int_0^x 1\, dx' \int_0^y 1\, dy' = xy. \end{aligned}$$

Practice Exercise 5.5. Let X and Y be two independent uniform random variables Gaussian(μ, σ^2). Find the joint CDF.

Solution. Let $\Phi(\cdot)$ be the CDF of the standard Gaussian.

$$\begin{aligned} F_{X,Y}(x,y) &= F_X(x)F_Y(y) \\ &= \int_{-\infty}^{x} f_X(x')\, dx' \int_{-\infty}^{y} f_Y(y')\, dy' = \Phi\left(\frac{x-\mu}{\sigma}\right) \Phi\left(\frac{y-\mu}{\sigma}\right). \end{aligned}$$

Here are a few properties of the CDF:

$$F_{X,Y}(x,-\infty) = \int_{-\infty}^{-\infty}\int_{-\infty}^{x} f_{X,Y}(x',y')\,dx'\,dy' = \int_{-\infty}^{x} 0\,dx' = 0,$$
$$F_{X,Y}(-\infty,y) = \int_{-\infty}^{y}\int_{-\infty}^{-\infty} f_{X,Y}(x',y')\,dx'\,dy' = \int_{-\infty}^{y} 0\,dy' = 0,$$
$$F_{X,Y}(-\infty,-\infty) = \int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty} f_{X,Y}(x',y')\,dx'\,dy' = 0,$$
$$F_{X,Y}(\infty,\infty) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x',y')\,dx'\,dy' = 1.$$

In addition, we can obtain the marginal CDF as follows.

Proposition 5.1. *Let X and Y be two random variables. The* **marginal CDF** *is*

$$F_X(x) = F_{X,Y}(x,\infty), \tag{5.10}$$
$$F_Y(y) = F_{X,Y}(\infty,y). \tag{5.11}$$

Proof. We prove only the first case. The second case is similar.

$$F_{X,Y}(x,\infty) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(x',y')\,dy'\,dx' = \int_{-\infty}^{y} f_X(x')\,dx' = F_X(x). \qquad \square$$

By the fundamental theorem of calculus, we can derive the PDF from the CDF.

Definition 5.9. *Let $F_{X,Y}(x,y)$ be the joint CDF of X and Y. Then, the joint PDF is*

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial y\,\partial x} F_{X,Y}(x,y). \tag{5.12}$$

The order of the partial derivatives can be switched, yielding a symmetric result:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\,\partial y} F_{X,Y}(x,y).$$

Example 5.14. Let X and Y be two uniform random variables with joint CDF $F_{X,Y}(x,y) = xy$ for $0 \le x \le 1$ and $0 \le y \le 1$. Find the joint PDF.

Solution.

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} xy = 1,$$

which is consistent with the definition of a joint uniform random variable.

Practice Exercise 5.6. Let X and Y be two exponential random variables with joint CDF

$$F_{X,Y}(x,y) = (1-e^{-\lambda x})(1-e^{-\lambda y}), \qquad x \geq 0,\ y \geq 0.$$

Find the joint PDF.

Solution.

$$\begin{aligned} f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) &= \frac{\partial^2}{\partial x \partial y}(1-e^{-\lambda x})(1-e^{-\lambda y}) \\ &= \frac{\partial}{\partial x}\left((1-e^{-\lambda x})(\lambda e^{-\lambda y})\right) = \lambda e^{-\lambda x}\lambda e^{-\lambda y}. \end{aligned}$$

which is consistent with the definition of a joint exponential random variable.

5.2 Joint Expectation

5.2.1 Definition and interpretation

When we have a single random variable, the expectation is defined as

$$\mathbb{E}[X] = \int_{\Omega} x f_X(x)\, dx.$$

For a pair of random variables, what would be a good way of defining the expectation? Certainly, we cannot just replace $f_X(x)$ by $f_{X,Y}(x,y)$ because the integration has to become a double integration. However, if it is a double integration, where should we put the variable y? It turns out that a useful way of defining the expectation for X and Y is as follows.

Definition 5.10. *Let X and Y be two random variables. The* **joint expectation** *is*

$$\mathbb{E}[XY] = \sum_{y \in \Omega_Y} \sum_{x \in \Omega_X} xy \cdot p_{X,Y}(x,y) \tag{5.13}$$

if X and Y are discrete, or

$$\mathbb{E}[XY] = \int_{y \in \Omega_Y} \int_{x \in \Omega_X} xy \cdot f_{X,Y}(x,y)\, dx\, dy \tag{5.14}$$

if X and Y are continuous. Joint expectation is also called **correlation**.

The double summation and integration on the right-hand side of the equation is nothing but the state times the probability. Here, the state is the product xy, and the probability is the joint PMF $p_{X,Y}(x,y)$ (or PDF). Therefore, as long as you agree that joint expectation should be defined as $\mathbb{E}[XY]$, the double summation and the double integration make sense.

The biggest mystery here is $\mathbb{E}[XY]$. You may wonder why the joint expectation should be defined as the expectation of the *product* $\mathbb{E}[XY]$. Why not the sum $\mathbb{E}[X+Y]$, or the difference $\mathbb{E}[X-Y]$, or the quotient $\mathbb{E}[X/Y]$? Why are we so deeply interested in X times Y? These are excellent questions. That the joint expectation is defined as the product has to do with the correlation between two random variables. We will take a small detour into linear algebra.

Let us consider two discrete random variables X and Y, both with N states. So X will take the states $\{x_1, x_2, \ldots, x_N\}$ and Y will take the states $\{y_1, y_2, \ldots, y_N\}$. Let's define them as two vectors: $\boldsymbol{x} \overset{\text{def}}{=} [x_1, \ldots, x_N]^T$ and $\boldsymbol{y} \overset{\text{def}}{=} [y_1, \ldots, y_N]^T$. Since X and Y are random variables, they have a joint PMF $p_{X,Y}(x, y)$. The array of the PMF values can be written as a matrix:

$$\text{PMF as a matrix} = \boldsymbol{P} \overset{\text{def}}{=} \begin{bmatrix} p_{X,Y}(x_1, y_1) & p_{X,Y}(x_1, y_2) & \cdots & p_{X,Y}(x_1, y_N) \\ p_{X,Y}(x_2, y_1) & p_{X,Y}(x_2, y_2) & \cdots & p_{X,Y}(x_2, y_N) \\ \vdots & \vdots & \ddots & \vdots \\ p_{X,Y}(x_N, y_1) & p_{X,Y}(x_N, y_2) & \cdots & p_{X,Y}(x_N, y_N) \end{bmatrix}.$$

Let's try to write the joint expectation in terms of matrices and vectors. The definition of a joint expectation tells us that

$$\mathbb{E}[XY] = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i y_j \cdot p_{X,Y}(x_i, y_j),$$

which can be written as

$$\mathbb{E}[XY] = \underbrace{\begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix}}_{\boldsymbol{x}^T} \underbrace{\begin{bmatrix} p_{X,Y}(x_1, y_1) & \cdots & p_{X,Y}(x_1, y_N) \\ \vdots & \ddots & \vdots \\ p_{X,Y}(x_N, y_1) & \cdots & p_{X,Y}(x_N, y_N) \end{bmatrix}}_{\boldsymbol{P}} \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}}_{\boldsymbol{y}} = \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{y}.$$

This is a *weighted* inner product between $\boldsymbol{x}$ and $\boldsymbol{y}$ using the weight matrix $\boldsymbol{P}$.

Why correlation is defined as $\mathbb{E}[XY]$

- $\mathbb{E}[XY]$ is a weighted inner product between the states:

$$\mathbb{E}[XY] = \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{y}.$$

- $\boldsymbol{x}$ and $\boldsymbol{y}$ are the states of the random variables X and Y.
- The inner product measures the similarity between two vectors.

Example 5.15. Let X be a discrete random variable with N states, where each state has an equal probability. Thus, $p_X(x) = 1/N$ for all x. Let $Y = X$ be another variable.

Then the joint PMF of (X, Y) is

$$p_{X,Y}(x, y) = \begin{cases} \frac{1}{N}, & x = y, \\ 0, & x \neq y. \end{cases}$$

It follows that the joint expectation is

$$\mathbb{E}[XY] = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i y_j \cdot p_{X,Y}(x_i, y_j) = \frac{1}{N} \sum_{i=1}^{N} x_i y_i.$$

Equivalently, we can obtain the result via the inner product by defining

$$\boldsymbol{P} = \begin{bmatrix} \frac{1}{N} & 0 & \cdots & 0 \\ 0 & \frac{1}{N} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \frac{1}{N} \end{bmatrix} = \frac{1}{N} \boldsymbol{I}.$$

In this case, the weighted inner product is

$$\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{y} = \frac{\boldsymbol{x}^T \boldsymbol{y}}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i y_i = \mathbb{E}[XY].$$

How do we understand the inner product? Ignoring the matrix $\boldsymbol{P}$ for a moment, we recall an elementary result in linear algebra.

Definition 5.11. *Let $\boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{y} \in \mathbb{R}^N$ be two vectors. Define the* **cosine angle** *$\cos\theta$ as*

$$\cos\theta = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}, \tag{5.15}$$

where $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{N} x_i^2}$ is the **norm** *of the vector $\boldsymbol{x}$, and $\|\boldsymbol{y}\| = \sqrt{\sum_{i=1}^{N} y_i^2}$ is the norm of the vector $\boldsymbol{y}$.*

This definition can be understood as the geometry between two vectors, as illustrated in **Figure 5.8**. If the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel so that $\boldsymbol{x} = \alpha \boldsymbol{y}$ for some α, then the angle $\theta = 0$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal so that $\boldsymbol{x}^T \boldsymbol{y} = 0$, then $\theta = \pi/2$. Therefore, the inner product $\boldsymbol{x}^T \boldsymbol{y}$ tells us the degree of correlation between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

Now let's come back to our discussion about the joint expectation. The cosine angle definition tells us that if $\mathbb{E}[XY] = \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{y}$, the following form would make sense:

$$\cos\theta = \frac{\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} = \frac{\mathbb{E}[XY]}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}.$$

That is, as long as we can find out the norms $\|\boldsymbol{x}\|$ and $\|\boldsymbol{y}\|$, we will be able to interpret $\mathbb{E}[XY]$ from the cosine angle perspective. But what would be a reasonable definition of $\|\boldsymbol{x}\|$

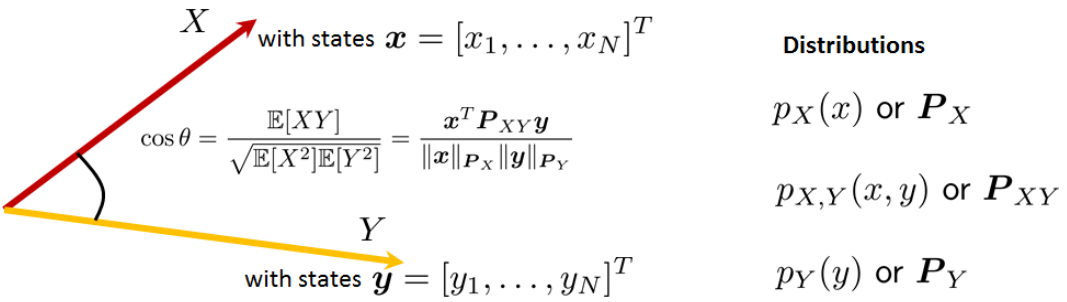


Figure 5.8: The geometry of joint expectation. $\mathbb{E}[XY]$ gives us the cosine angle between the two random variables. This, in turn, tells us the correlation between the two random variables.

and $\|\boldsymbol{y}\|$? We define the norm by first considering the variance of the random variable X and Y:

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_{i=1}^{N} x_i x_i \cdot p_X(x_i) \\
&= \underbrace{\begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix}}_{\boldsymbol{x}^T} \underbrace{\begin{bmatrix} p_X(x_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_X(x_N) \end{bmatrix}}_{\boldsymbol{P}_X} \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}}_{\boldsymbol{x}} \\
&= \boldsymbol{x}^T \boldsymbol{P}_X \boldsymbol{x} = \|\boldsymbol{x}\|^2_{\boldsymbol{P}_X},
\end{aligned}$$

where $\boldsymbol{P}_X$ is the diagonal matrix storing the probability masses of the random variable X. It is not difficult to show that $\boldsymbol{P}_X = \text{diag}(\boldsymbol{P}\boldsymbol{1})$ by following the definition of the marginal distributions (which are the column and row sums of the joint PMF). Similarly we can define

$$\begin{aligned}
\mathbb{E}[Y^2] &= \sum_{j=1}^{N} y_j y_j \cdot p_Y(y_j) \\
&= \underbrace{\begin{bmatrix} y_1 & \cdots & y_N \end{bmatrix}}_{\boldsymbol{y}^T} \underbrace{\begin{bmatrix} p_Y(y_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_Y(y_N) \end{bmatrix}}_{\boldsymbol{P}_Y} \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}}_{\boldsymbol{y}} \\
&= \boldsymbol{y}^T \boldsymbol{P}_Y \boldsymbol{y} = \|\boldsymbol{y}\|^2_{\boldsymbol{P}_Y}.
\end{aligned}$$

Therefore, one way to define the cosine angle is to start with

$$\cos\theta = \frac{\boldsymbol{x}^T \boldsymbol{P}_{XY} \boldsymbol{y}}{\|\boldsymbol{x}\|_{\boldsymbol{P}_X} \|\boldsymbol{y}\|_{\boldsymbol{P}_Y}},$$

where $\boldsymbol{P}_{XY} = \boldsymbol{P}$, $\|\boldsymbol{x}\|_{\boldsymbol{P}_X} = \sqrt{\boldsymbol{x}^T \boldsymbol{P}_X \boldsymbol{x}}$ and $\|\boldsymbol{y}\|_{\boldsymbol{P}_Y} = \sqrt{\boldsymbol{y}^T \boldsymbol{P}_Y \boldsymbol{y}}$. But writing it in terms of the expectation, we observe that this cosine angle is exactly

$$\cos\theta = \frac{\boldsymbol{x}^T \boldsymbol{P}_{XY} \boldsymbol{y}}{\|\boldsymbol{x}\|_{\boldsymbol{P}_X} \|\boldsymbol{y}\|_{\boldsymbol{P}_Y}} = \frac{\mathbb{E}[XY]}{\sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}}.$$

Therefore, $\mathbb{E}[XY]$ defines the cosine angle between the two random variables, which, in turn, defines the correlation between the two. A large $|\mathbb{E}[XY]|$ means that X and Y are highly correlated, and a small $|\mathbb{E}[XY]|$ means that X and Y are not very correlated. If $\mathbb{E}[XY] = 0$, then the two random variables are uncorrelated. Therefore, $\mathbb{E}[XY]$ tells us how the two random variables are related to each other.

To further convince you that $\frac{\mathbb{E}[XY]}{\sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}}$ can be interpreted as a cosine angle, we show that

$$-1 \leq \frac{\mathbb{E}[XY]}{\sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}} \leq 1,$$

because if this ratio can go beyond $+1$ and -1, it makes no sense to call it a cosine angle. The argument follows from a very well-known inequality in probability, called the Cauchy-Schwarz inequality (for expectation), which states that $-1 \leq \frac{\mathbb{E}[XY]}{\sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}} \leq 1$:

Theorem 5.2 (Cauchy-Schwarz inequality). *For any random variables X and Y,*

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]. \tag{5.16}$$

The following proof can be skipped if you are reading the book the first time.

Proof. Let $t \in \mathbb{R}$ be a constant. Consider $\mathbb{E}[(X + tY)^2] = \mathbb{E}[X^2 + 2tXY + t^2Y^2]$. Since $\mathbb{E}[(X + tY)^2] \geq 0$ for any t, it follows that

$$\mathbb{E}[X^2 + 2tXY + t^2Y^2] \geq 0.$$

Expanding the left-hand side yields $t^2\mathbb{E}[Y^2] + 2t\mathbb{E}[XY] + \mathbb{E}[X^2] \geq 0$. This is a quadratic equation in t, and we know that for any quadratic equation $at^2 + bt + c \geq 0$ we must have $b^2 - 4ac \leq 0$. Therefore, in our case, we have that

$$(2\mathbb{E}[XY])^2 - 4\mathbb{E}[Y^2]\mathbb{E}[X^2] \leq 0,$$

which means $(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$. The equality holds when $\mathbb{E}[(X + tY)^2] = 0$. In this case, $X = -tY$ for some t, i.e., the random variable X is a scaled version of Y so that the vector formed by the states of X is parallel to that of Y.

□

End of the proof.

5.2.2 Covariance and correlation coefficient

In many practical problems, we prefer to work with central moments, i.e., $\mathbb{E}[(X - \mu_X)^2]$ instead of $\mathbb{E}[X^2]$. This essentially means that we subtract the mean from the random variable. If we adopt such a centralized random variable, we can define the **covariance** as follows.

Definition 5.12. *Let X and Y be two random variables. Then the* **covariance** *of X and Y is*

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)], \tag{5.17}$$

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

It is easy to show that if $X = Y$, then the covariance simplifies to the variance:

$$\begin{aligned}\text{Cov}(X, X) &= \mathbb{E}[(X - \mu_X)(X - \mu_X)] \\ &= \text{Var}[X].\end{aligned}$$

Thus, covariance is a generalization of variance. The former can handle a pair of variables, whereas the latter is only for a single variable. We can also demonstrate the following result.

Theorem 5.3. *Let X and Y be two random variables. Then*

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{5.18}$$

Proof. Just apply the definition of covariance:

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= \mathbb{E}[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] \\ &= \mathbb{E}[XY] - \mu_X\mu_Y.\end{aligned}$$

□

The next theorem concerns the sum of two random variables.

Theorem 5.4. *For any X and Y,*

a. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

b. $\text{Var}[X + Y] = \text{Var}[X] + 2\text{Cov}(X, Y) + \text{Var}[Y]$.

Proof. Recall the definition of joint expectation:

$$\begin{aligned}\mathbb{E}[X + Y] &= \sum_y \sum_x (x + y) p_{X,Y}(x, y) \\ &= \sum_y \sum_x x p_{X,Y}(x, y) + \sum_y \sum_x y p_{X,Y}(x, y) \\ &= \sum_x x \left(\sum_y p_{X,Y}(x, y) \right) + \sum_y y \left(\sum_x p_{X,Y}(x, y) \right) \\ &= \sum_x x p_X(x) + \sum_y y p_Y(y) \\ &= \mathbb{E}[X] + \mathbb{E}[Y].\end{aligned}$$

Similarly,

$$\begin{aligned}\mathrm{Var}[X+Y] &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 \\ &= \mathbb{E}[(X+Y)^2] - (\mu_X+\mu_Y)^2 \\ &= \mathbb{E}[X^2+2XY+Y^2] - (\mu_X^2+2\mu_X\mu_Y+\mu_Y^2) \\ &= \mathbb{E}[X^2] - \mu_X^2 + \mathbb{E}[Y^2] - \mu_Y^2 + 2(\mathbb{E}[XY]-\mu_X\mu_Y) \\ &= \mathrm{Var}[X] + 2\mathrm{Cov}(X,Y) + \mathrm{Var}[Y].\end{aligned}$$

□

With covariance defined, we can now define the **correlation coefficient** ρ, which is the cosine angle of the centralized variables. That is,

$$\begin{aligned}\rho &= \cos\theta \\ &= \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\mathbb{E}[(X-\mu_X)^2]\mathbb{E}[(Y-\mu_Y)^2]}}.\end{aligned}$$

Recognizing that the denominator of this expression is just the variance of X and Y, we define the correlation coefficient as follows.

Definition 5.13. *Let X and Y be two random variables. The* **correlation coefficient** *is*

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}. \tag{5.19}$$

Since $-1 \leq \cos\theta \leq 1$, ρ is also between -1 and 1. The difference between ρ and $\mathbb{E}[XY]$ is that ρ is *normalized* with respect to the variance of X and Y, whereas $\mathbb{E}[XY]$ is not normalized. The correlation coefficient has the following properties:

- ρ is always between -1 and 1, i.e., $-1 \leq \rho \leq 1$. This is due to the cosine angle definition.
- When $X = Y$ (fully correlated), $\rho = +1$.
- When $X = -Y$ (negatively correlated), $\rho = -1$.
- When X and Y are uncorrelated, $\rho = 0$.

5.2.3 Independence and correlation

If two random variables X and Y are independent, the joint expectation can be written as a product of two individual expectations.

Theorem 5.5. *If X and Y are independent, then*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]. \tag{5.20}$$

Proof. We only prove the discrete case because the continuous can be proved similarly. If X and Y are independent, we have $p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$. Therefore,

$$\begin{aligned}\mathbb{E}[XY] = \sum_y \sum_x xy p_{X,Y}(x,y) &= \sum_y \sum_x xy p_X(x) p_Y(y) \\ &= \left(\sum_x x p_X(x)\right)\left(\sum_y y p_Y(y)\right) = \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

□

In general, for any two independent random variables and two functions f and g,

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

The following theorem illustrates a few important relationships between independence and correlation.

Theorem 5.6. *Consider the following two statements:*

a. X *and* Y *are independent;*

b. $\mathrm{Cov}(X,Y) = 0$.

Statement (a) implies statement (b), but (b) does not imply (a). Thus, independence is a stronger condition than correlation.

Proof. We first prove that (a) implies (b). If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. In this case,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

To prove that (b) does not imply (a), we show a counterexample. Consider a discrete random variable Z with PMF

$$p_Z(z) = \begin{bmatrix}\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4}\end{bmatrix}.$$

Let X and Y be

$$X = \cos\frac{\pi}{2}Z \quad \text{and} \quad Y = \sin\frac{\pi}{2}Z.$$

Then we can show that $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y] = 0$. The covariance is

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[(X-0)(Y-0)] \\ &= \mathbb{E}\left[\cos\frac{\pi}{2}Z \sin\frac{\pi}{2}Z\right] \\ &= \mathbb{E}\left[\frac{1}{2}\sin\pi Z\right] \\ &= \frac{1}{2}\left[(\sin\pi 0)\frac{1}{4} + (\sin\pi 1)\frac{1}{4} + (\sin\pi 2)\frac{1}{4} + (\sin\pi 3)\frac{1}{4}\right] = 0.\end{aligned}$$

The next step is to show that X and Y are dependent. To this end, we only need to show that $p_{X,Y}(x,y) \neq p_X(x)p_Y(y)$. The joint PMF $p_{X,Y}(x,y)$ can be found by noting that

$$\begin{aligned} Z = 0 &\Rightarrow X = 1,\ Y = 0, \\ Z = 1 &\Rightarrow X = 0,\ Y = 1, \\ Z = 2 &\Rightarrow X = -1,\ Y = 0, \\ Z = 3 &\Rightarrow X = 0,\ Y = -1. \end{aligned}$$

Thus, the PMF is

$$p_{X,Y}(x,y) = \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 \end{bmatrix}.$$

The marginal PMFs are

$$p_X(x) = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}, \quad p_Y(y) = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}.$$

The product $p_X(x)\, p_Y(y)$ is

$$p_X(x)p_Y(y) = \begin{bmatrix} \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \end{bmatrix}.$$

Therefore, $p_{X,Y}(x,y) \neq p_X(x)p_Y(y)$, although $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

□

What is the relationship between independent and uncorrelated?

- Independent $\Rightarrow$ uncorrelated.
- Independent $\not\Leftarrow$ uncorrelated.

5.2.4 Computing correlation from data

We close this section by discussing a very practical problem: Given a dataset containing two columns of data points, how do we determine whether the two columns are correlated?

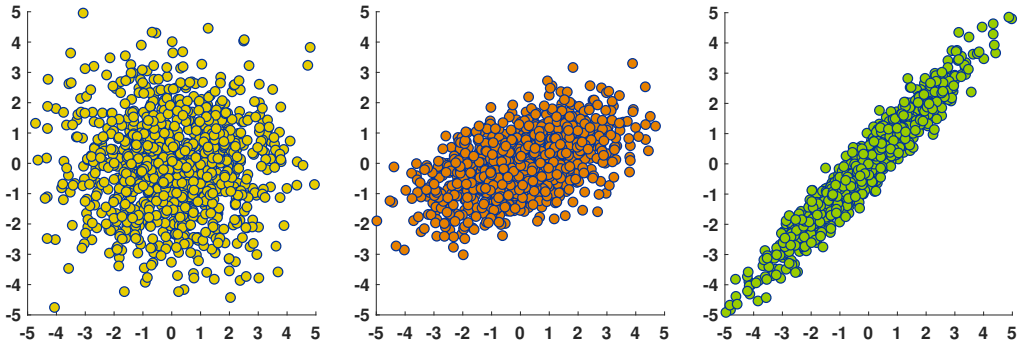
Recall that the correlation coefficient is defined as

$$\rho = \frac{\mathbb{E}[XY] - \mu_X \mu_Y}{\sigma_X \sigma_Y}.$$

If we have a dataset containing $(x_n, y_i)_{n=1}^N$, then the correlation coefficient can be approximated by

$$\widehat{\rho} = \frac{\frac{1}{N}\sum_{n=1}^N x_n y_n - \overline{x}\,\overline{y}}{\sqrt{\frac{1}{N}\sum_{n=1}^N (x_n - \overline{x})^2}\sqrt{\frac{1}{N}\sum_{n=1}^N (y_n - \overline{y})^2}},$$

where $\overline{x} = \frac{1}{N}\sum_{n=1}^N x_n$ and $\overline{y} = \frac{1}{N}\sum_{n=1}^N y_n$ are the means. This equation should not be a surprise because essentially all terms are the empirical estimates. Thus, $\widehat{\rho}$ is the empirical correlation coefficient determined from the dataset. As $N \to \infty$, we expect $\widehat{\rho} \to \rho$.



(a) $\widehat{\rho} = -0.0038$ (b) $\widehat{\rho} = 0.5321$ (c) $\widehat{\rho} = 0.9656$

Figure 5.9: Visualization of correlated variables. Each of these figures represent a scattered plot of a dataset containing $(x_n, y_n)_{n=1}^N$. (a) is uncorrelated. (b) is somewhat correlated. (c) is strongly correlated.

Figure 5.9 shows three example datasets. We plot the (x_n, y_n) pairs as coordinates in the 2D plane. The first dataset contains samples that are almost uncorrelated. We can see that x_n does not tell us anything about y_n. The second dataset is moderately correlated. The third dataset is highly correlated: If we know x_n, we are almost certain to know the corresponding y_n, with a small number of perturbations.

On a computer, computing the correlation coefficient can be done using built-in commands such as `corrcoef` in MATLAB and `stats.pearsonr` in Python. The codes to generate the results in **Figure 5.9**(b) are shown below.

```
% MATLAB code to compute the correlation coefficient
x = mvnrnd([0,0],[3 1; 1 1],1000);
figure(1); scatter(x(:,1),x(:,2));
rho = corrcoef(x)
```

```
# Python code to compute the correlation coefficient
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt
x = stats.multivariate_normal.rvs([0,0], [[3,1],[1,1]], 10000)
plt.figure(); plt.scatter(x[:,0],x[:,1])
rho,_ = stats.pearsonr(x[:,0],x[:,1])
print(rho)
```

5.3 Conditional PMF and PDF

Whenever we have a pair of random variables X and Y that are correlated, we can define their conditional distributions, which quantify the probability of $X = x$ given $Y = y$. In this section, we discuss the concepts of conditional PMF and PDF.

5.3.1 Conditional PMF

We start by defining the conditional PMF for a pair of discrete random variables.

Definition 5.14. *Let X and Y be two discrete random variables. The* **conditional PMF** *of X given Y is*

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}. \tag{5.21}$$

The simplest way to understand this is to view $p_{X|Y}(x|y)$ as $\mathbb{P}[X = x \,|\, Y = y]$. That is, given that $Y = y$, what is the probability for $X = x$? To see why this perspective makes sense, let us recall the definition of a conditional probability:

$$\begin{aligned} p_{X|Y}(x|y) &= \frac{p_{X,Y}(x,y)}{p_Y(y)} \\ &= \frac{\mathbb{P}[X = x \cap Y = y]}{\mathbb{P}[Y = y]} = \mathbb{P}[X = x \,|\, Y = y]. \end{aligned}$$

As we can see, the last two equalities are essentially the definitions of conditional probability and the joint PMF.

How should we understand the notation $p_{X|Y}(x|y)$? Is it a one-variable function in x or a two-variable function in (x, y)? What does $p_{X|Y}(x|y)$ tell us? To answer these questions, let us first try to understand the randomness exhibited in a conditional PMF. In $p_{X|Y}(x|y)$, the random variable Y is *fixed* to a specific value $Y = y$. Therefore there is nothing random about Y. All the possibilities of Y have already been taken care of by the denominator $p_Y(y)$. Only the variable x in $p_{X|Y}(x|y)$ has randomness. What do we mean by "fixed at a value $Y = y$"? Consider the following example.

Example 5.16. Suppose there are two coins. Let

$$\begin{aligned} X &= \text{the sum of the values of two coins,} \\ Y &= \text{the value of the first coin.} \end{aligned}$$

Clearly, X has 3 states: 0, 1, 2, and Y has two states: either 0 or 1. When we say $p_{X|Y}(x|1)$, we refer to the probability mass function of X when fixing $Y = 1$. If we do not impose this condition, the probability mass of X is simple:

$$p_X(x) = \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right].$$

However, if we include the conditioning, then

$$\begin{aligned} p_{X|Y}(x|1) &= \frac{p_{X,Y}(x,1)}{p_Y(1)} \\ &= \frac{\left[0, \frac{2}{4}, \frac{1}{4}\right]}{\frac{1}{6}} = \left[0, \frac{2}{3}, \frac{1}{3}\right]. \end{aligned}$$

To put this in plain words, when $Y = 1$, there is no way for X to take the state 0. The chance for X to take the state 1 is $2/3$ because either $(0, 1)$ or $(1, 0)$ can give $X = 1$. The chance for X to take the state 2 is $1/3$ because it has to be $(1, 1)$ in order to give $X = 2$. Therefore, when we say "conditioned on $Y = 1$", we mean that we limit our observations to cases where $Y = 1$. Since Y is already fixed at $Y = 1$, there is nothing random about Y. The only variable is X. This example is illustrated in **Figure 5.10**.

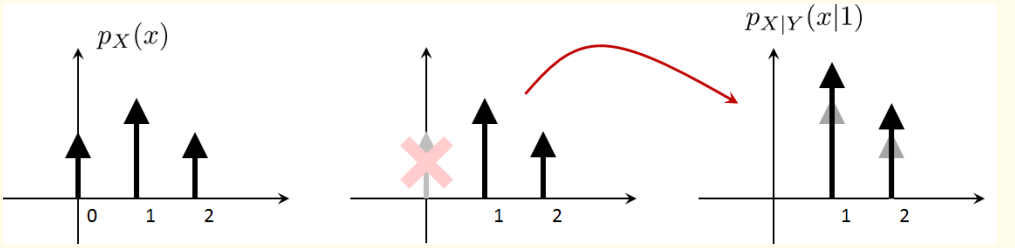


Figure 5.10: Suppose X is the sum of two coins with PMF $0.25, 0.5, 0.25$. Let Y be the first coin. When X is unconditioned, the PMF is just $[0.25, 0.5, 0.25]$. When X is conditioned on $Y = 1$, then "$X = 0$" cannot happen. Therefore, the resulting PMF $p_{X|Y}(x|1)$ only has two states. After normalization we obtain the conditional PMF $[0, 0.66, 0.33]$.

Since Y is already fixed at a particular value $Y = y$, $p_{X|Y}(x|y)$ is a probability mass function of x (we want to emphasize again that it is x and not y). So $p_{X|Y}(x|y)$ is a one-variable function in x. It is not the same as the usual PMF $p_X(x)$. $p_{X|Y}(x|y)$ is conditioned on $Y = y$. For example, $p_{X|Y}(x|1)$ is the PMF of X restricted to the condition that $Y = 1$. In fact, it follows that

$$\sum_{x\in\Omega_X} p_{X|Y}(x|y) = \sum_{x\in\Omega_X} \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{\sum_{x\in\Omega_X} p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_Y(y)}{p_Y(y)} = 1,$$

but this tells us that $p_{X|Y}(x|y)$ is a legitimate probability mass of X. If we sum over the y's instead, then we will hit a bump:

$$\sum_{y\in\Omega_Y} p_{X|Y}(x|y) = \sum_{y\in\Omega_Y} \frac{p_{X,Y}(x,y)}{p_Y(y)} \neq 1.$$

Therefore, while $p_{X|Y}(x|y)$ is a legitimate probability mass function of X, it is not a probability mass function of Y.

Example 5.17. Consider a joint PMF given in the following table. Find the conditional PMF $p_{X|Y}(x|1)$ and the marginal PMF $p_X(x)$.

	Y=			
	1	2	3	4
X = 1	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{0}{20}$
2	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{1}{20}$
3	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{1}{20}$
4	$\frac{0}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$

Solution. To find the marginal PMF, we sum over all the y's for every x:

$$x=1: \quad p_X(1)=\sum_{y=1}^{4} p_{X,Y}(1,y)=\frac{1}{20}+\frac{1}{20}+\frac{1}{20}+\frac{0}{20}=\frac{3}{20},$$

$$x=2: \quad p_X(2)=\sum_{y=1}^{4} p_{X,Y}(2,y)=\frac{1}{20}+\frac{2}{20}+\frac{2}{20}+\frac{1}{20}=\frac{6}{20},$$

$$x=3: \quad p_X(3)=\sum_{y=1}^{4} p_{X,Y}(3,y)=\frac{1}{20}+\frac{3}{20}+\frac{3}{20}+\frac{1}{20}=\frac{8}{20},$$

$$x=4: \quad p_X(4)=\sum_{y=1}^{4} p_{X,Y}(4,y)=\frac{0}{20}+\frac{1}{20}+\frac{1}{20}+\frac{1}{20}=\frac{3}{20}.$$

Hence, the marginal PMF is

$$p_X(x)=\begin{bmatrix}\frac{3}{20} & \frac{6}{20} & \frac{8}{20} & \frac{3}{20}\end{bmatrix}.$$

The conditional PMF $p_{X|Y}(x|1)$ is

$$p_{X|Y}(x|1)=\frac{p_{X,Y}(x,1)}{p_Y(1)}=\frac{\begin{bmatrix}\frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{0}{20}\end{bmatrix}}{\frac{3}{20}}=\begin{bmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0\end{bmatrix}.$$

Practice Exercise 5.7. Consider two random variables X and Y defined as follows.

$$Y=\begin{cases}10^2, & \text{with prob } 5/6,\\ 10^4, & \text{with prob } 1/6.\end{cases} \qquad X=\begin{cases}10^{-4}Y, & \text{with prob } 1/2,\\ 10^{-3}Y, & \text{with prob } 1/3,\\ 10^{-2}Y, & \text{with prob } 1/6.\end{cases}$$

Find $p_{X|Y}(x|y)$, $p_X(x)$ and $p_{X,Y}(x,y)$.

Solution. Since Y takes two different states, we can enumerate $Y=10^2$ and $Y=10^4$. This gives us

$$p_{X|Y}(x|10^2)=\begin{cases}1/2, & \text{if } x=0.01,\\ 1/3, & \text{if } x=0.1,\\ 1/6, & \text{if } x=1.\end{cases}$$

$$p_{X|Y}(x|10^4)=\begin{cases}1/2, & \text{if } x=1,\\ 1/3, & \text{if } x=10,\\ 1/6, & \text{if } x=100.\end{cases}$$

The joint PMF $p_{X,Y}(x,y)$ is

$$p_{X,Y}(x,10^2) = p_{X|Y}(x|10^2)p_Y(10^2) = \begin{cases} \left(\frac{1}{2}\right)\left(\frac{5}{6}\right), & x = 0.01, \\ \left(\frac{1}{3}\right)\left(\frac{5}{6}\right), & x = 0.1, \\ \left(\frac{1}{6}\right)\left(\frac{5}{6}\right), & x = 1. \end{cases}$$

$$p_{X,Y}(x,10^4) = p_{X|Y}(x|10^4)p_Y(10^4) = \begin{cases} \left(\frac{1}{2}\right)\left(\frac{1}{6}\right), & x = 1, \\ \left(\frac{1}{3}\right)\left(\frac{1}{6}\right), & x = 10, \\ \left(\frac{1}{6}\right)\left(\frac{1}{6}\right), & x = 100. \end{cases}$$

Therefore, the joint PMF is given by the following table.

10^4	0	0	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$
10^2	$\frac{5}{12}$	$\frac{5}{18}$	$\frac{5}{36}$	0	0
	0.01	0.1	1	10	100

The marginal PMF $p_X(x)$ is thus

$$p_X(x) = \sum_y p_{X,Y}(x,y) = \begin{bmatrix} \frac{5}{12} & \frac{5}{18} & \frac{2}{9} & \frac{1}{18} & \frac{1}{36} \end{bmatrix}.$$

In the previous two examples, what is the probability $\mathbb{P}[X \in A \,|\, Y = y]$ or the probability $\mathbb{P}[X \in A]$ for some events A? The answers are giving by the following theorem.

Theorem 5.7. *Let X and Y be two discrete random variables, and let A be an event. Then*

$$(i) \qquad \mathbb{P}[X \in A \,|\, Y = y] = \sum_{x \in A} p_{X|Y}(x|y)$$

$$(ii) \qquad \mathbb{P}[X \in A] = \sum_{x \in A} \sum_{y \in \Omega_Y} p_{X|Y}(x|y)p_Y(y) = \sum_{y \in \Omega_Y} \mathbb{P}[X \in A \,|\, Y = y]p_Y(y).$$

Proof. The first statement is based on the fact that if A contains a finite number of elements, then $\mathbb{P}[X \in A]$ is equivalent to the sum $\sum_{x \in A} \mathbb{P}[X = x]$. Thus,

$$\begin{aligned} \mathbb{P}[X \in A \,|\, Y = y] &= \frac{\mathbb{P}[X \in A \cap Y = y]}{\mathbb{P}[Y = y]} \\ &= \frac{\sum_{x \in A} \mathbb{P}[X = x \cap Y = y]}{\mathbb{P}[Y = y]} \\ &= \sum_{x \in A} p_{X|Y}(x|y). \end{aligned}$$

The second statement holds because the inner summation $\sum_{y \in \Omega_Y} p_{X|Y}(x|y)p_Y(y)$ is just the marginal PMF $p_X(x)$. Thus the outer summation yields the probability.

□

Example 5.18. Let us follow up on Example 5.17. What is the probability that $\mathbb{P}[X > 2|Y = 1]$? What is the probability that $\mathbb{P}[X > 2]$?

Solution. Since the problem asks about the conditional probability, we know that it can be computed by using the conditional PMF. This gives us

$$\begin{aligned}\mathbb{P}[X > 2|Y = 1] &= \sum_{x>2} p_{X|Y}(x|1) \\ &= \cancel{p_{X|Y}(1|1)} + \cancel{p_{X|Y}(2|1)} + \underbrace{p_{X|Y}(3|1)}_{\frac{1}{3}} + \underbrace{p_{X|Y}(4|1)}_{0} = \frac{1}{3}.\end{aligned}$$

The other probability is

$$\begin{aligned}\mathbb{P}[X > 2] &= \sum_{x>2} p_X(x) \\ &= \cancel{p_X(1)} + \cancel{p_X(2)} + \underbrace{p_X(3)}_{\frac{8}{20}} + \underbrace{p_X(4)}_{\frac{3}{20}} = \frac{11}{20}.\end{aligned}$$

What is the rule of thumb for conditional distribution?

- The PMF/PDF should *match* with the probability you are finding.
- If you want to find the conditional probability $\mathbb{P}[X \in A|Y = y]$, use the conditional PMF $p_{X|Y}(x|y)$.
- If you want to find the probability $\mathbb{P}[X \in A]$, use the marginal PMF $p_X(x)$.

Finally, we define the conditional CDF for discrete random variables.

Definition 5.15. *Let X and Y be discrete random variables. Then the* **conditional CDF** *of X given $Y = y$ is*

$$F_{X|Y}(x|y) = \mathbb{P}[X \leq x \,|\, Y = y] = \sum_{x' \leq x} p_{X|Y}(x'|y). \tag{5.22}$$

5.3.2 Conditional PDF

We now discuss the conditioning of a continuous random variable.

Definition 5.16. *Let X and Y be two continuous random variables. The* **conditional PDF** *of X given Y is*

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}. \tag{5.23}$$

Example 5.19. Let X and Y be two continuous random variables with a joint PDF

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x}e^{-y}, & 0 \le y \le x < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

Find the conditional PDFs $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$.

Solution. We first find the marginal PDFs.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_0^x 2e^{-x}e^{-y}\,dy = 2e^{-x}(1-e^{-x}),$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_y^{\infty} 2e^{-x}e^{-y}\,dx = 2e^{-2y}.$$

Thus, the conditional PDFs are

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{2e^{-x}e^{-y}}{2e^{-2y}} = e^{-(x+y)}, \quad x \ge y, \\ f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \frac{2e^{-x}e^{-y}}{2e^{-x}(1-e^{-x})} = \frac{e^{-y}}{1-e^{-x}}, \quad 0 \le y < x. \end{aligned}$$

Where does the conditional PDF come from? We cannot duplicate the argument we used for the discrete case because the denominator of a conditional PMF becomes $\mathbb{P}[Y = y] = 0$ when Y is continuous. To answer this question, we first define the conditional CDF for continuous random variables.

Definition 5.17. *Let X and Y be continuous random variables. Then the* **conditional CDF** *of X given $Y = y$ is*

$$F_{X|Y}(x|y) = \frac{\int_{-\infty}^{x} f_{X,Y}(x',y)\,dx'}{f_Y(y)}. \tag{5.24}$$

Why should the conditional CDF of continuous random variable be defined in this way? One way to interpret $F_{X|Y}(x|y)$ is as the limiting perspective. We can define the **conditional CDF** as

$$\begin{aligned} F_{X|Y}(x|y) &= \lim_{h\to 0} \mathbb{P}(X \le x \,|\, y \le Y \le y+h) \\ &= \lim_{h\to 0} \frac{\mathbb{P}(X \le x \cap y \le Y \le y+h)}{\mathbb{P}[y \le Y \le y+h]}. \end{aligned}$$

With some calculations, we have that

$$\begin{aligned}\lim_{h\to 0}\frac{\mathbb{P}(X\le x\cap y\le Y\le y+h)}{\mathbb{P}[y\le Y\le y+h]} &= \lim_{h\to 0}\frac{\int_{-\infty}^{x}\int_{y}^{y+h} f_{X,Y}(x',y')\,dy'\,dx'}{\int_{y}^{y+h} f_Y(y')\,dy'}\\ &= \lim_{h\to 0}\frac{\int_{-\infty}^{x} f_{X,Y}(x',y')\,dx'\cdot h}{f_Y(y)\cdot h}\\ &= \frac{\int_{-\infty}^{x} f_{X,Y}(x',y')\,dx'}{f_Y(y)}.\end{aligned}$$

The key here is that the small step size h in the numerator and the denominator will cancel each other out. Now, given the conditional CDF, we can verify the definition of the conditional PDF. It holds that

$$\begin{aligned}f_{X|Y}(x|y) &= \frac{d}{dx}F_{X|Y}(x|y)\\ &= \frac{d}{dx}\left\{\frac{\int_{-\infty}^{x} f_{X,Y}(x',y)\,dx'}{f_Y(y)}\right\}\overset{(a)}{=}\frac{f_{X,Y}(x,y)}{f_Y(y)},\end{aligned}$$

where (a) follows from the fundamental theorem of calculus.

Just like the conditional PMF, we can calculate the probabilities using the conditional PDFs. In particular, if we evaluate the probability where $X\in A$ given that Y takes a particular value $Y=y$, then we can integrate the conditional PDF $f_{X|Y}(x|y)$, with respect to x.

Theorem 5.8. *Let X and Y be continuous random variables, and let A be an event.*

(i) $\mathbb{P}[X\in A\mid Y=y]=\int_A f_{X|Y}(x|y)\,dx$,

(ii) $\mathbb{P}[X\in A]=\int_{\Omega_Y}\mathbb{P}[X\in A\mid Y=y]f_Y(y)\,dy$.

Example 5.20. Let X be a random bit such that

$$X=\begin{cases}+1, & \text{with prob } 1/2,\\ -1, & \text{with prob } 1/2.\end{cases}$$

Suppose that X is transmitted over a noisy channel so that the observed signal is

$$Y=X+N,$$

where $N\sim\text{Gaussian}(0,1)$ is the noise, which is independent of the signal X. Find the probabilities $\mathbb{P}[X=+1\mid Y>0]$ and $\mathbb{P}[X=-1\mid Y>0]$.

Solution. First, we know that

$$f_{Y|X}(y|+1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(y-1)^2}{2}}\qquad\text{and}\qquad f_{Y|X}(y|-1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(y+1)^2}{2}}.$$

Therefore, integrating y from 0 to ∞ gives us

$$\begin{aligned}\mathbb{P}[Y > 0 \,|\, X = +1] &= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}}\, dy \\ &= 1 - \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}}\, dy \\ &= 1 - \Phi\left(\frac{0-1}{1}\right) = 1 - \Phi(-1).\end{aligned}$$

Similarly, we have $\mathbb{P}[Y > 0 \,|\, X = -1] = 1 - \Phi(+1)$. The probability we want to find is $\mathbb{P}[X = +1 \,|\, Y > 0]$, which can be determined using Bayes' theorem.

$$\mathbb{P}[X = +1 \,|\, Y > 0] = \frac{\mathbb{P}[Y > 0 \,|\, X = +1]\mathbb{P}[X = +1]}{\mathbb{P}[Y > 0]}.$$

The denominator can be found by using the law of total probability:

$$\begin{aligned}\mathbb{P}[Y > 0] &= \mathbb{P}[Y > 0 \,|\, X = +1]\mathbb{P}[X = +1] \\ &\quad + \mathbb{P}[Y > 0 \,|\, X = -1]\mathbb{P}[X = -1] \\ &= 1 - \frac{1}{2}\left(\Phi(+1) + \Phi(-1)\right) \\ &= \frac{1}{2},\end{aligned}$$

since $\Phi(+1) + \Phi(-1) = \Phi(+1) + 1 - \Phi(+1) = 1$. Therefore,

$$\begin{aligned}\mathbb{P}[X = +1 \,|\, Y > 0] &= 1 - \Phi(-1) \\ &= 0.8413.\end{aligned}$$

The implication is that if $Y > 0$, the probability $\mathbb{P}[X = +1 \,|\, Y > 0] = 0.8413$. The complement of this result gives $\mathbb{P}[X = -1 \,|\, Y > 0] = 1 - 0.8413 = 0.1587$.

Practice Exercise 5.8. Find $\mathbb{P}[Y > y]$, where

$$X \sim \text{Uniform}[1, 2], \quad Y \,|\, X \sim \text{Exponential}(x).$$

Solution. The tricky part of this problem is the tendency to confuse the two variables X and Y. Once you understand their roles the problem becomes easy. First notice that $Y \,|\, X \sim \text{Exponential}(x)$ is a conditional distribution. It says that given $X = x$, the probability distribution of Y is exponential, with the parameter x. Thus, we have that

$$f_{Y|X}(y|x) = xe^{-xy}.$$

Why? Recall that if $Y \sim \text{Exponential}(\lambda)$ then $f_Y(y) = \lambda e^{-\lambda y}$. Now if we replace λ with x, we have xe^{-xy}. So the role of x in this conditional density function is as a parameter.

Stanley H. Chan is an associate professor of electrical and computer engineering, and an associate professor of statistics, at Purdue University, West Lafayette. His research areas include computational photography, image processing, and machine learning. At Purdue, he teaches undergraduates probability and graduates machine learning. He is a recipient of Purdue University College of Engineering Exceptional Early Career Teaching Award, the Ruth and Joel Spira Outstanding Teaching Award, Purdue Teaching for Tomorrow Fellow, among other awards.