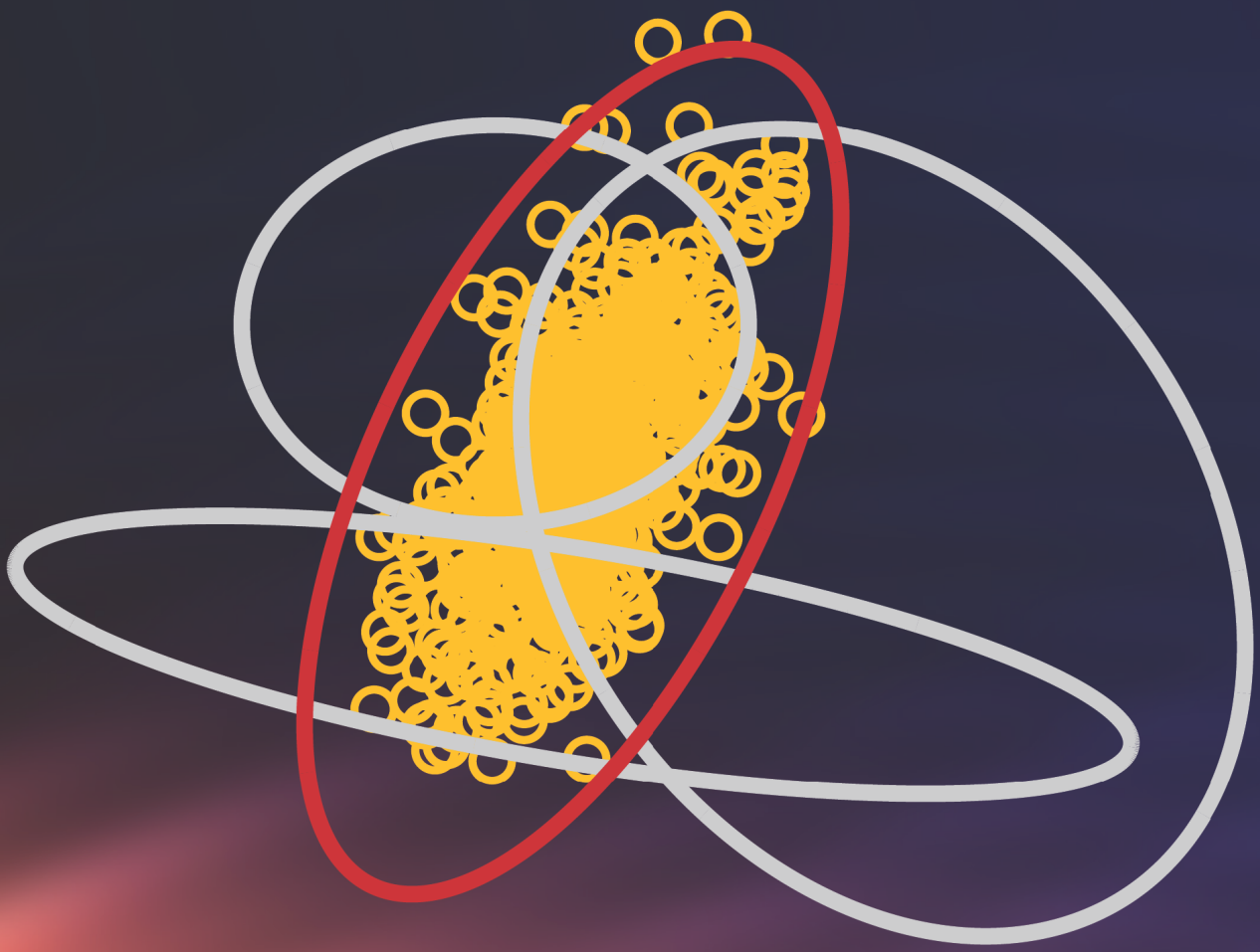
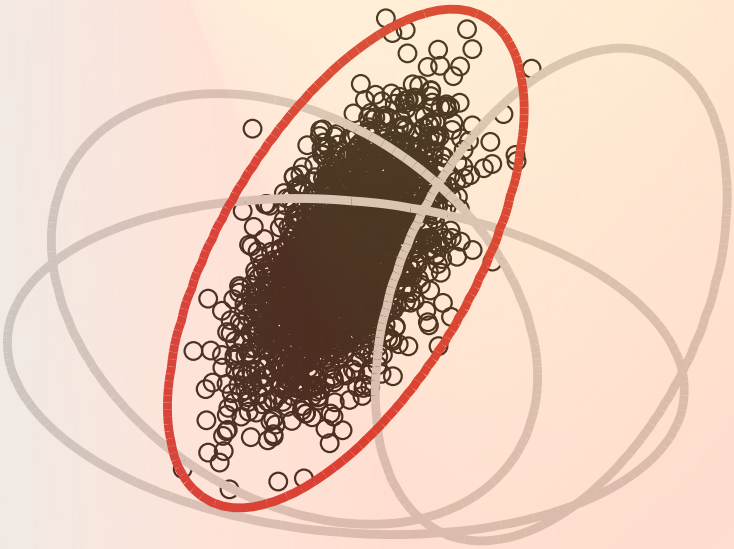
Introduction to PROBABILITY for DATA SCIENCE



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 $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$, 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.

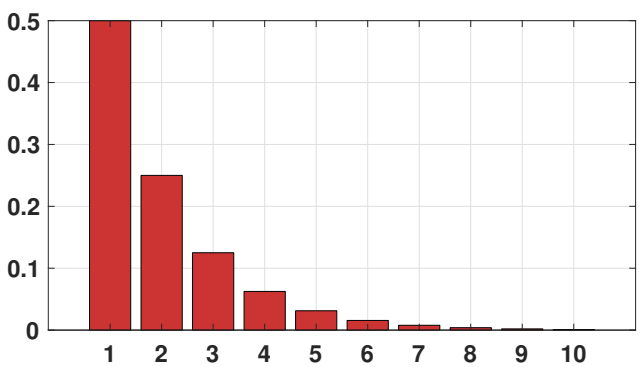


Figure 1.2: The histogram of flipping a coin until we see a head. The x-axis is the number of coin flips, and the y-axis is the probability.

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*

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

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

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

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

□

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:

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

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 $k \le n$. 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.

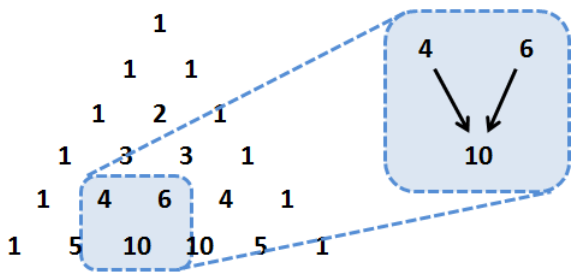


Figure 1.4: Pascal triangle for $n = 0, \ldots, 5$. Note that a number in one row is obtained by summing two numbers directly above it.

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.

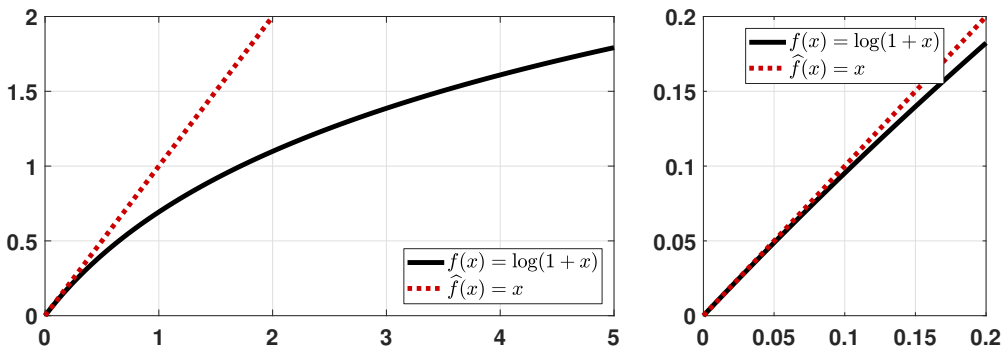


Figure 1.5: The function $f(x) = \log(1+x)$ and the approximation $\widehat{f}(x) = x$.

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

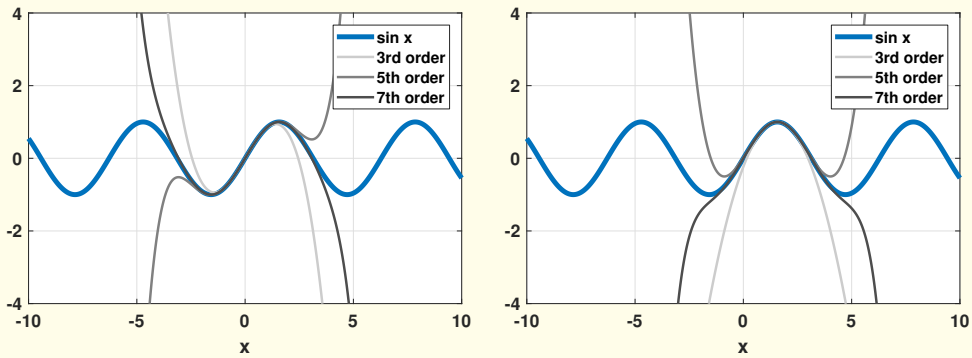
$$f(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

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**.

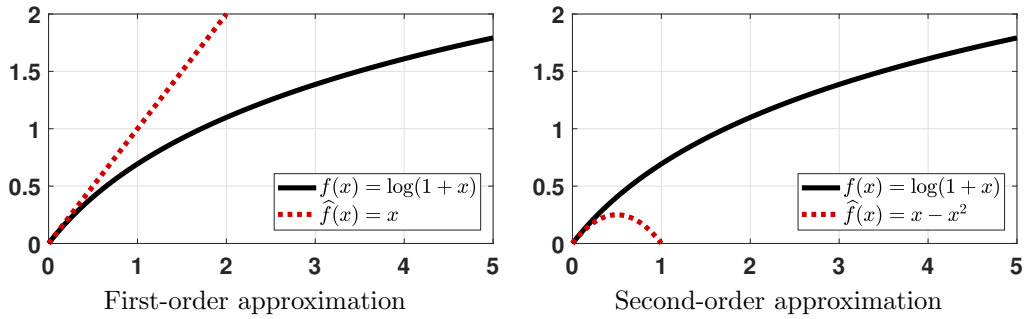


Figure 1.7: The function $f(x) = \log(1+x)$, the first-order approximation $\widehat{f}(x) = x$, and the second-order approximation $\widehat{f}(x) = x - x^2$.

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.

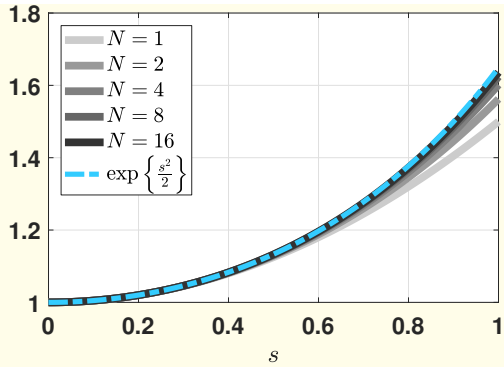


Figure 1.8: We plot a sequence of function $f_N(x) = \left(1 + \frac{s^2}{2N}\right)^N$ and its limit $f(x) = e^{s^2/2}$.

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

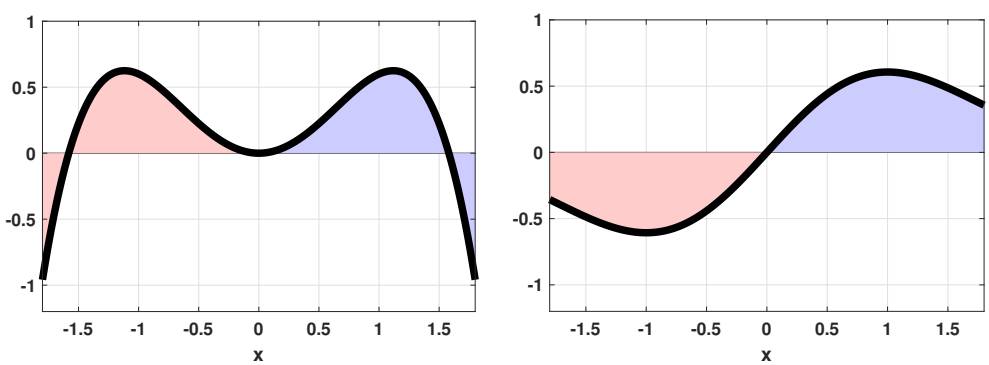
$$\int_{-1}^{1} f(x)\ dx = 2\int_{0}^{1} f(x)\ dx = 2\int_{0}^{1} x^2 - 0.4^4\ dx = 2\left[\frac{x^3}{3} - \frac{0.4}{5}x^5\right]_{x=0}^{x=1} = \frac{38}{75}.$$

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 $F(2)$.

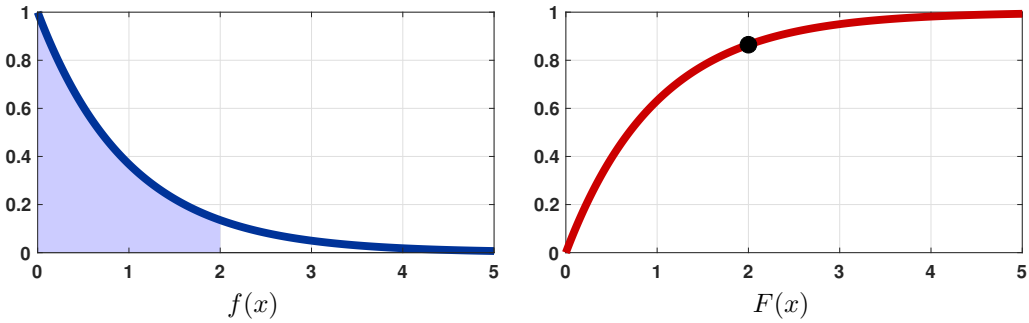


Figure 1.10: The pair of functions $f(x) = e^{-x}$ and $F(x) = 1 - e^{-x}$

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,

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

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 $\mathbf{0} = [0, \ldots, 0]^T$.

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.

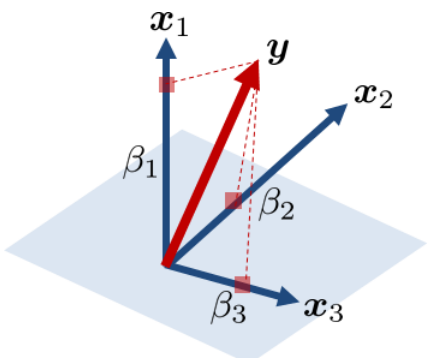


Figure 1.11: Representing an observable vector $\boldsymbol{y}$ by a linear combination of variable vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$. The combination weights are $\beta_1, \beta_2, \beta_3$.

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
$\vdots$	$\vdots$
0.7379	3.4054
0.2691	2.5672
0.4228	2.3796
0.6020	3.2942

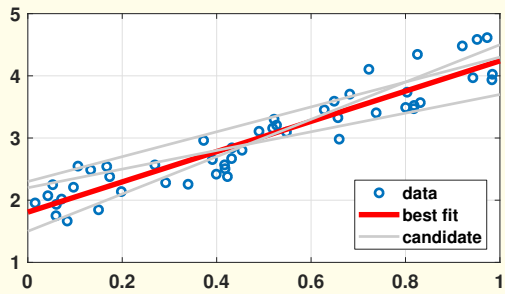


Figure 1.12: Example of fitting a set of data points. The problem can be described by $\boldsymbol{y} = \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2$.

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.

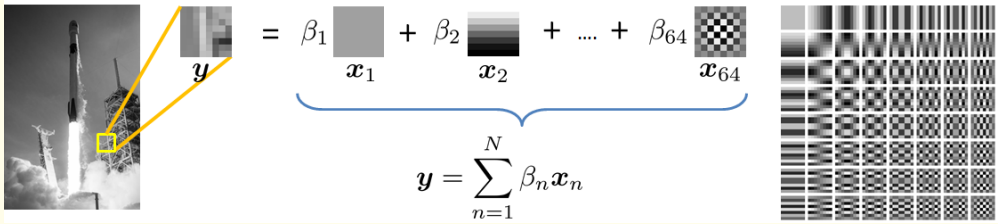


Figure 1.13: JPEG image compression is based on the concept of discrete cosine transform, which can be formulated as a matrix-vector problem.

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.

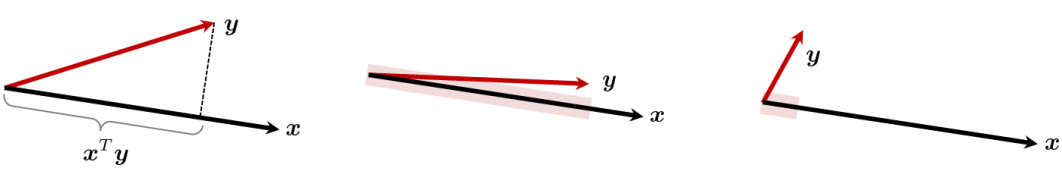


Figure 1.14: Geometric interpretation of inner product: We project one vector onto the other vector. The projected distance is the inner product.

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**.

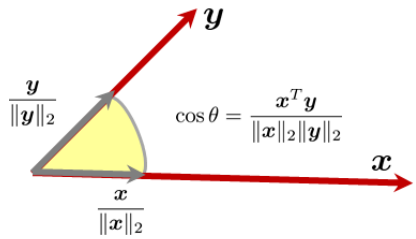


Figure 1.15: The cosine angle is the inner product divided by the norms of the vectors.

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

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

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?

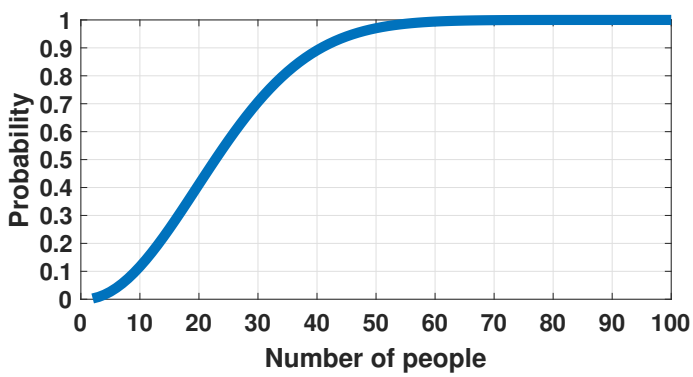


Figure 1.16: The probability for two people in a group to have the same birthday as a function of the number of people in the group.

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**.

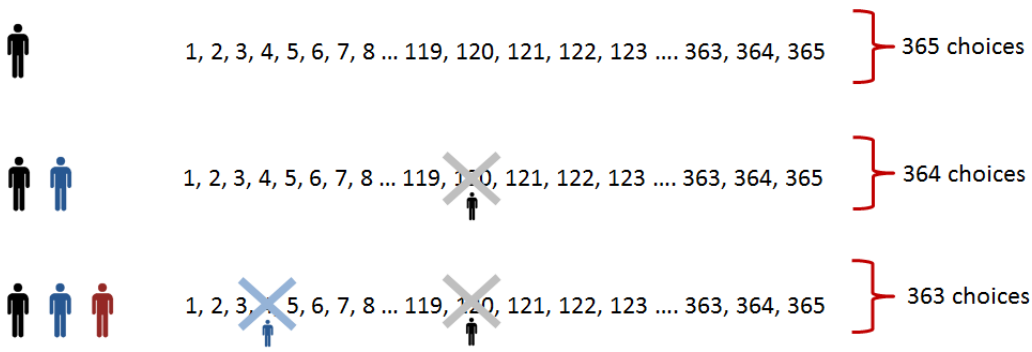


Figure 1.17: The probability for two people to have the same birthday as a function of the number of people in the group. When there is only one person, this person can land on any of the 365 days. When there are two people, the first person has already taken one day (out of 365 days), so the second 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**.

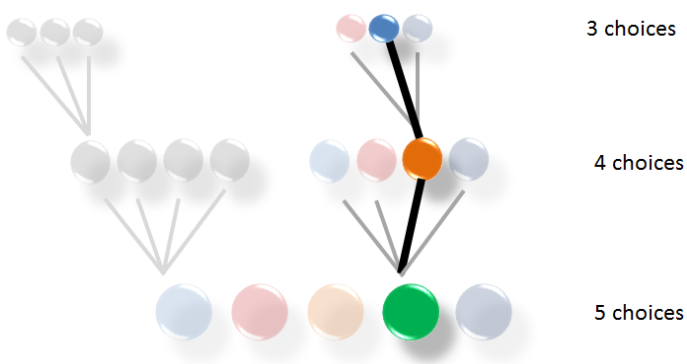


Figure 1.18: Permutation. The number of choices is reduced in every stage. Therefore, the total number is $n \times (n-1) \times \cdots \times (n-k+1)$ if there are k stages.

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.

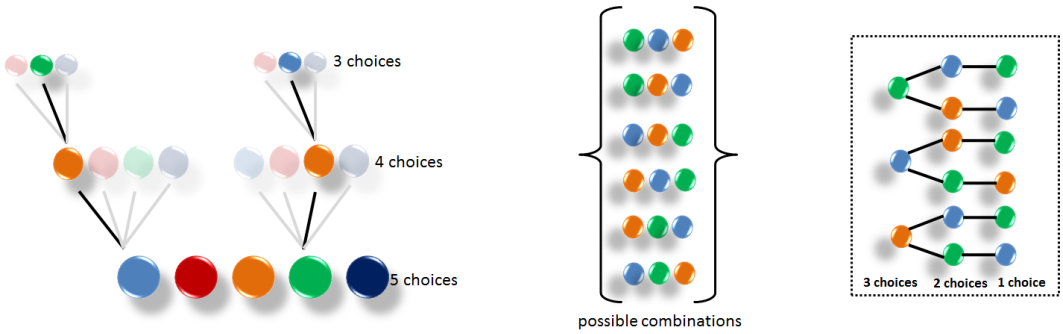


Figure 1.19: Combination. In this problem, we are interested in picking 3 colored balls out of 5. This will give us $5 \times 4 \times 3 = 60$ permutations. However, since we are not interested in the ordering, some of the permutations are repeated. For example, there are 6 combos of (green, light blue, 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

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

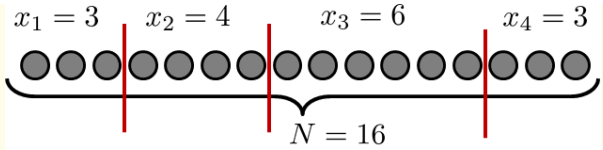


Figure 1.20: One possible solution for $N = 16$ and $K = 4$. In general, the problem is equivalent to inserting $K - 1$ dividers among $N - 1$ balls.

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

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

Exercise 2. (Video Solution)
Recall that

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

Evaluate

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

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.

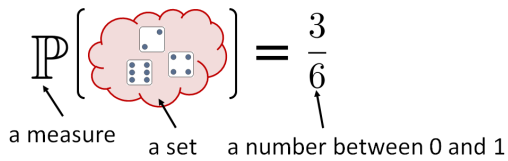


Figure 2.1: Probability is a measure of the size of a set. Whenever we talk about probability, it has to be the probability of a **set**.

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 $\mathbb{R}$.

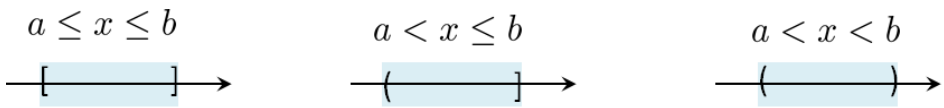


Figure 2.2: From left to right: a closed interval, a semi-closed (or semi-open) interval, and an open interval.

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).

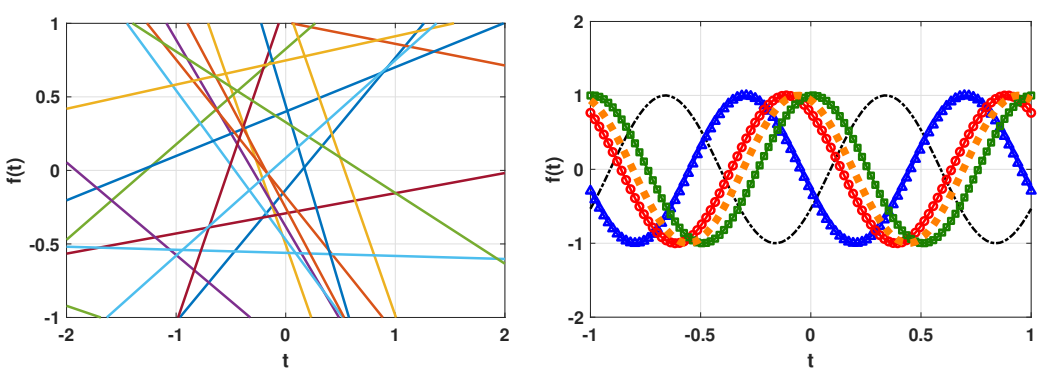


Figure 2.3: (a) The set of straight lines $A = \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = ax + b,\ a, b \in \mathbb{R}\}$. (b) The set of phase-shifted cosines $A = \{f : \mathbb{R} \to [-1, 1] \mid f(t) = \cos(\omega_0 t + \theta),\ \theta \in [0, 2\pi]\}$.

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 $A \cup B \neq \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = ax + b\}$ 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 $A \cup B = \Omega$.

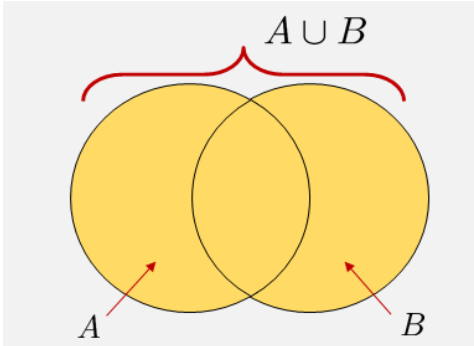


Figure 2.4: The union of two sets contains elements that are either in A or B or both.

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*

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

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

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

Similarly, an infinite union means that

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

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.

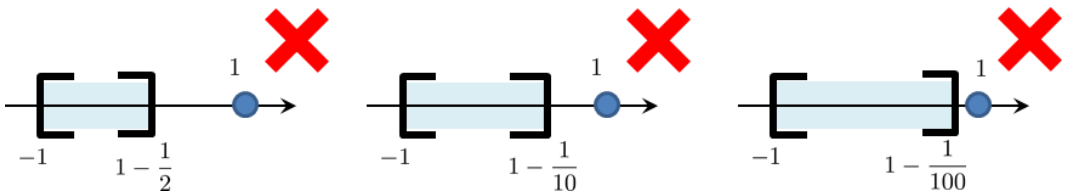


Figure 2.5: The infinite union of $\bigcup_{n=1}^{\infty} \left[-1, 1 - \frac{1}{n}\right]$. No matter how large n gets, the point 1 is never included. So the infinite union is $[-1, 1)$

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,*

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

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 $A \cap B \subseteq B$.

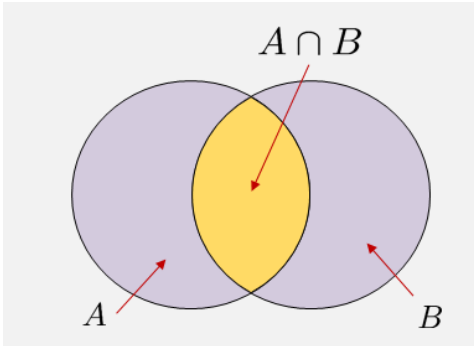


Figure 2.6: The intersection of two sets contains elements in both A and B.

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 $A = \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = ax\}$ and $B = \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = b\}$, 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, $A \cap B = \{f \mid f(x) = 0\}$.

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*

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

To understand this definition, we note that

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

As a result, it follows that

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

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

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

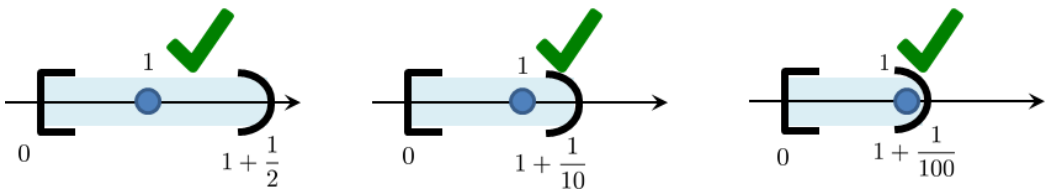


Figure 2.7: The infinite intersection of $\bigcap_{n=1}^{\infty} \left[0, 1 + \frac{1}{n}\right)$. No matter how large n gets, the point 1 is never included. So the infinite intersection is $[0, 1]$

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.

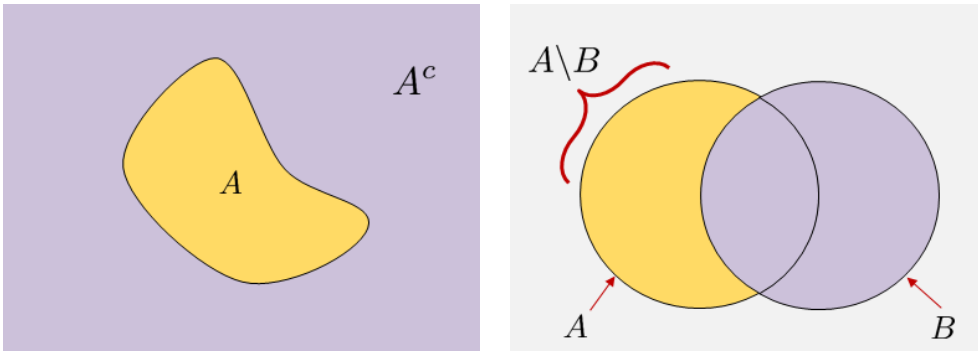


Figure 2.8: [Left] The complement of a set A contains all elements that are not in A. [Right] The difference $A\backslash B$ contains elements that are in A but not in B.

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.

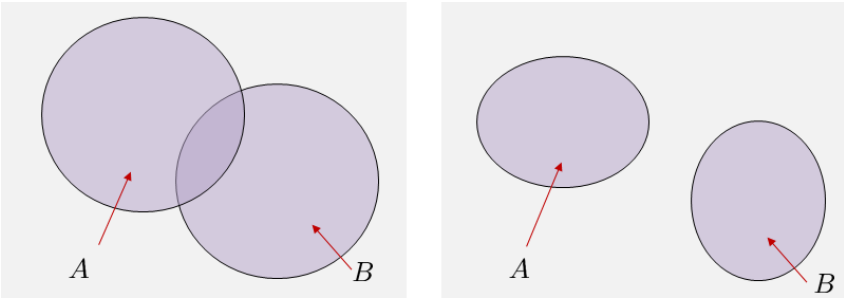


Figure 2.9: [Left] A and B are overlapping. [Right] A and B are disjoint.

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.

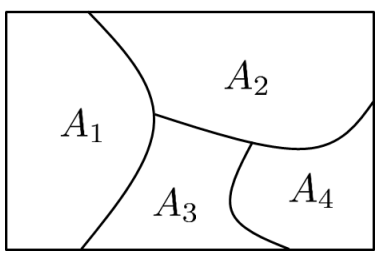


Figure 2.10: A partition of Ω contains disjoint subsets of which the union gives us Ω.

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)*

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

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

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

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.

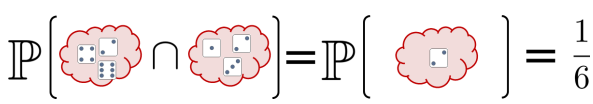


Figure 2.11: When there are two events A and B, the probability of $A \cap B$ is determined by first taking the intersection of the two sets and then evaluating its probability.

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**.

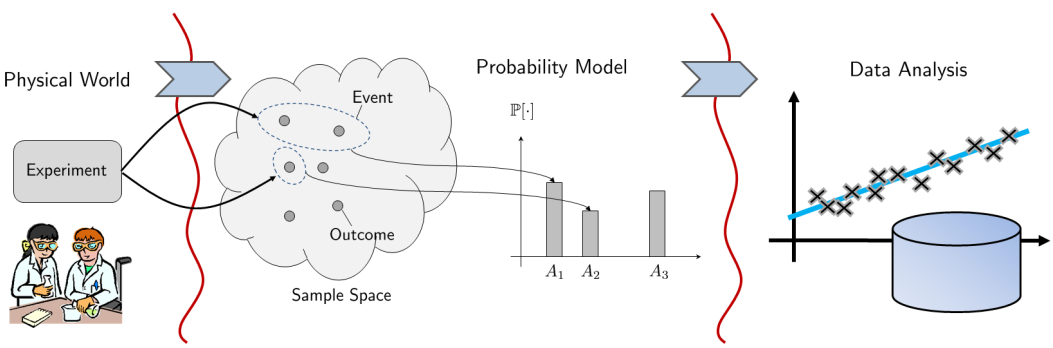


Figure 2.12: Given an experiment, we define the collection of all outcomes as the sample space. A subset in the sample space is called an event. The probability law is a mapping that maps an event to a number that denotes the size of the event.

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: $\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$.
- 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

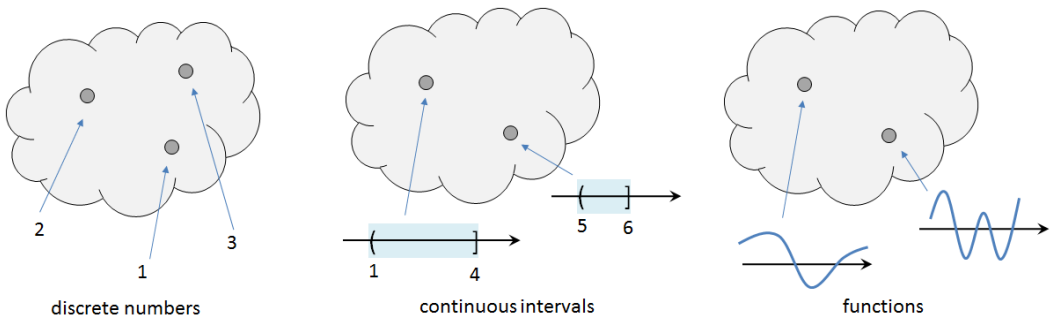


Figure 2.13: The sample space can take various forms: it can contain discrete numbers, or continuous intervals, or even functions.

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 {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}. 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 $\Omega =$ {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}. The following are two possible events, as illustrated in **Figure 2.14**.

- $E_1 = \{\text{even numbers}\} =$ {⚁, ⚃, ⚅}.

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

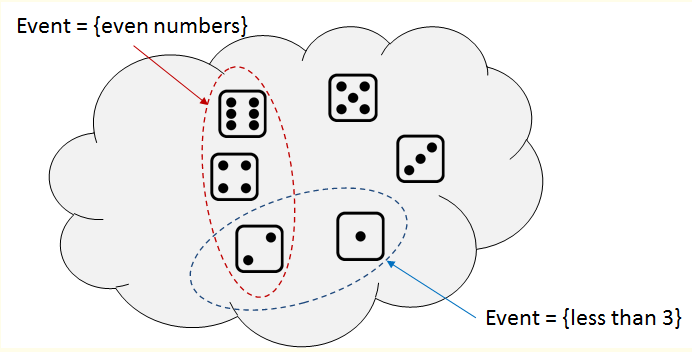


Figure 2.14: Two examples of events: The first event contains numbers $\{2, 4, 6\}$, and the second event contains numbers $\{1, 2\}$.

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.

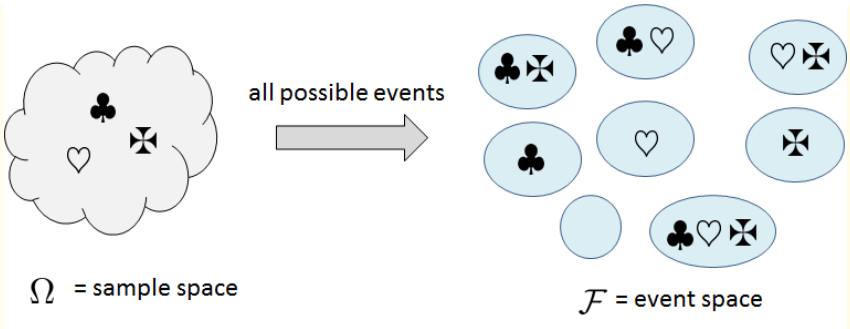


Figure 2.15: The event space contains all the possible subsets inside the sample space.

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 {⚀, ⚅} is encoded as the binary string 100010 because only ⚀ and ⚅ 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	⚀	⚁	⚂	⚃	⚄	⚅	Binary Code
$\emptyset$	×	×	×	×	×	×	000000
{⚀, ⚅}	○	×	×	×	○	×	100010
{⚂, ⚃, ⚄}	×	×	○	○	○	×	001110
⋮		⋮		⋮			⋮
{⚁, ⚂, ⚃, ⚄, ⚅}	×	○	○	○	○	○	011111
{⚀, ⚁, ⚂, ⚃, ⚄, ⚅}	○	○	○	○	○	○	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 $E_1 = \{⚀\}$ and $E_2 = \{⚁\}$ are events, it is necessary that $E = E_1 \cup E_2 = \{⚀, ⚁\}$ is an event too. Another example: if $E_1 = \{⚅, ⚅\}$ and $E_2 = \{⚀, ⚅\}$ are events, then it is necessary that $E = E_1 \cap E_2 = \{⚅\}$ is also an event. The third example: if $E_1 = \{⚂, ⚃, ⚅, ⚅\}$ is an event, then $E = E_1^c = \{⚀, ⚁\}$ 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 $E = \{⚂, ⚃, ⚅, ⚅\}$ is inside $\mathcal{F}$, then $E^c = \{⚀, ⚁\}$ 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 $E = \{⚂, ⚃, ⚅, ⚅\}$ (which is $E^c = \{⚀, ⚁\}$) 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.

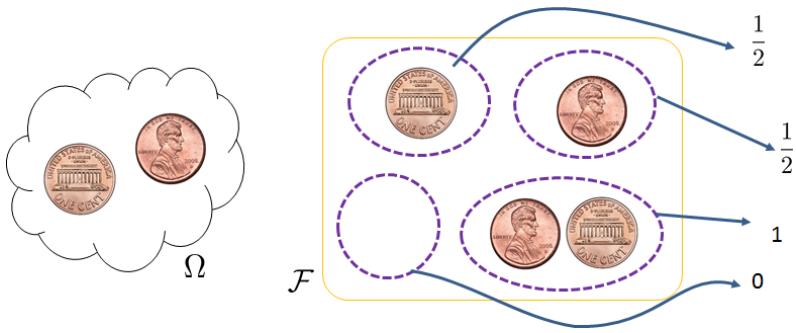


Figure 2.16: A probability law is a mapping from an event to a number. A probability law cannot be arbitrarily assigned; it must satisfy the axioms of probability.

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**.

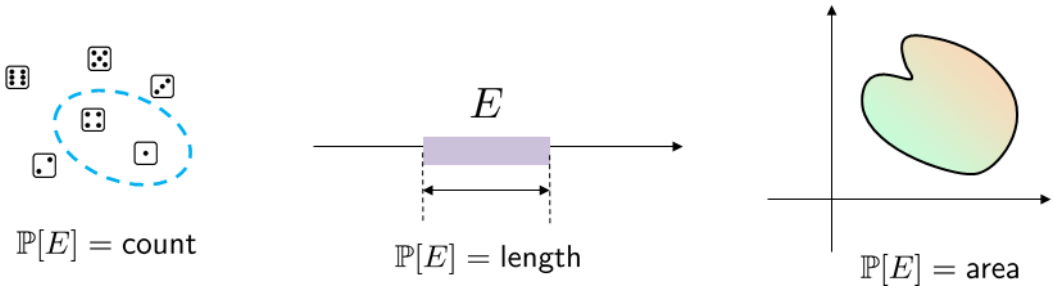


Figure 2.17: Probability is a measure of the size of a set. The probability can be a counter that counts the number of elements, a ruler that measures the length of an interval, or an integration that measures the area of a region.

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

$$\Omega = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\}.$$

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 $E = \{⚀, ⚂\}$ 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 $\Omega = \{♣, ♡, ✠\}$, and if $E_1 = \{♣\}$ and $E_2 = \{♣, ♡\}$, then one possible $\mathbb{P}$ is to assign $\mathbb{P}[E_1] = \mathbb{P}[\{♣\}] = \frac{1}{3}$ and $\mathbb{P}[E_2] = \mathbb{P}[\{♣, ♡\}] = 2/3$.

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 $\Omega = \{\clubsuit, \heartsuit, \maltese\}$. Let us assign a different probability to each outcome:

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

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 $\mathbb{P}[\{\heartsuit, \maltese\}]$ we find the sum of their weights $\mathbb{P}[\{\heartsuit, \maltese\}] = \frac{1}{6} + \frac{3}{6} = \frac{4}{6}$. Therefore, the notion of size does not refer to the number of elements but to the total weight of these elements.

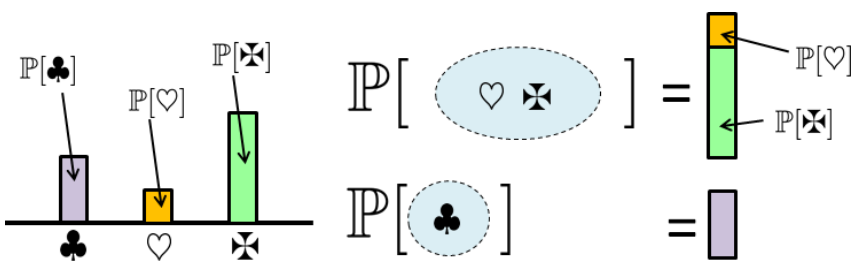


Figure 2.18: This example shows the "weights" of three elements in a set. The weights are numbers between 0 and 1 such that the sum is 1. When applying a probability measure to this set, we sum the weights for the elements in the events being considered. For example, $\mathbb{P}[\heartsuit, \maltese]$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

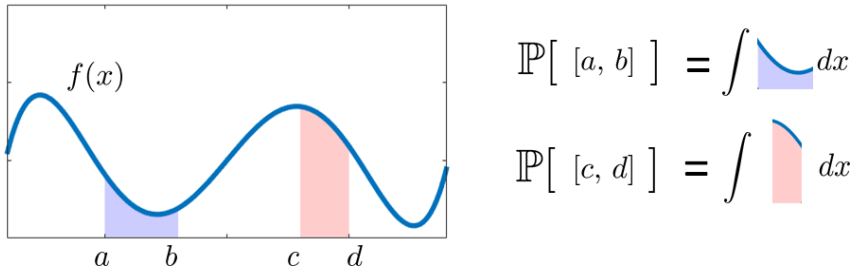


Figure 2.19: If the sample space is an interval on the real line, then the probability of an event is the area under the curve of the weighting function.

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.

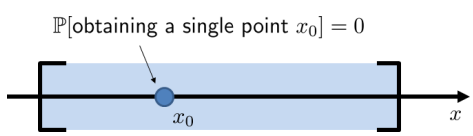


Figure 2.20: The probability of obtaining a single point in a continuous interval is zero.

[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.

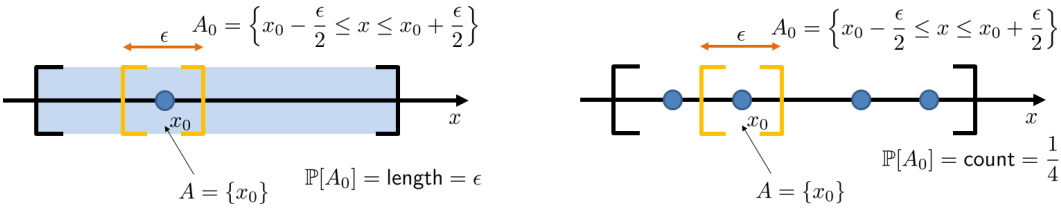


Figure 2.21: [Left] For a continuous sample space, a single point event $\{x_0\}$ can always be surrounded by a neighborhood A_0 whose size $\mathbb{P}[A_0] < \epsilon$. [Right] If you change the sample space to discrete elements, then a single point event $\{x_0\}$ can still be surrounded by a neighborhood 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 $\Omega = \{f : \mathbb{R} \to [-1, 1] \mid f(t) = \cos(\omega_0 t + \theta)\}$, where ω_0 is a fixed constant and θ is random. Construct a measure zero event and an almost sure event.

Solution. Let

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

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.

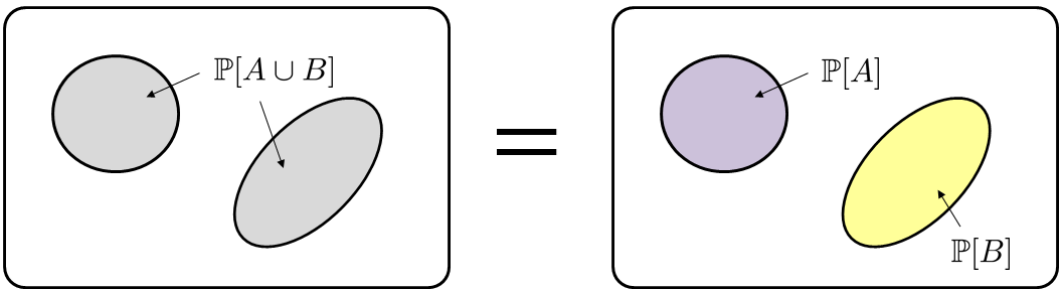


Figure 2.22: Axiom III says $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$ if $A \cap B = \emptyset$.

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 $\Omega = \{\clubsuit, \heartsuit, ✠\}$. The probability for each outcome is

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

Suppose we construct two disjoint events $E_1 = \{\clubsuit, \heartsuit\}$ and $E_2 = \{✠\}$. 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

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

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

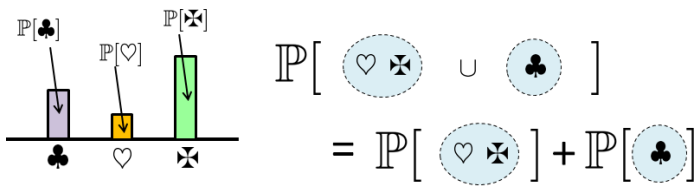


Figure 2.23: Applying weighting functions to the measures: Suppose we have three elements in the set. To compute the probability $\mathbb{P}[\{\heartsuit, ✠\} \cup \{\clubsuit\}]$, we can write it as the sum of $\mathbb{P}[\{\heartsuit, ✠\}]$ and $\mathbb{P}[\{\clubsuit\}]$.

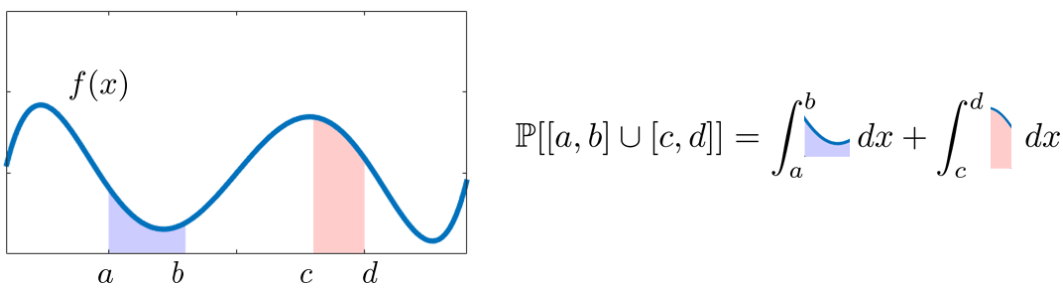


Figure 2.24: The axioms are compatible with the measure, even if we use a weighting function.

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.

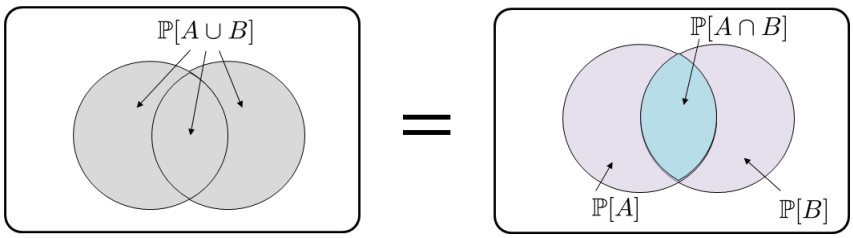


Figure 2.25: For any A and B, $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$.

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 $\Omega = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\}$ be the sample space of a fair die. Let $A = \{⚀, ⚁, ⚂\}$ and $B = \{⚂, ⚃, ⚄\}$. Then

$$\mathbb{P}[A \cup B] = \mathbb{P}[\{⚀, ⚁, ⚂, ⚃, ⚄\}] = \frac{5}{6}.$$

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

$$\Omega = \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = ax, \text{for all } a \in \mathbb{R}, x \in \mathbb{R}\}.$$

There are two events: $A = \{f \mid f(x) = ax,\ a \geq 0\}$, and $B = \{f \mid f(x) = ax,\ a \leq 0\}$. 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.

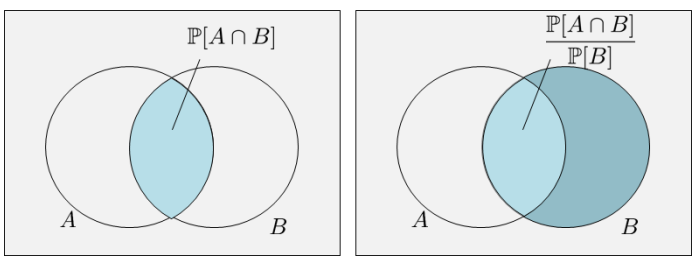


Figure 2.26: Illustration of conditional probability and its comparison with $\mathbb{P}[A \cap B]$.

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

$$\begin{aligned} A &= \{\text{Purdue gets Big Ten championship}\}, \\ B &= \{\text{Purdue wins 15 games consecutively}\}. \end{aligned}$$

In this example,

$$\begin{aligned} \mathbb{P}[A] &= \text{Prob. that Purdue gets the championship}, \\ \mathbb{P}[B] &= \text{Prob. that Purdue wins 15 games consecutively}, \\ \mathbb{P}[A \cap B] &= \text{Prob. that Purdue gets the championship and wins 15 games}, \\ \mathbb{P}[A \mid B] &= \text{Prob. that Purdue gets the championship given that} \\ &\qquad \text{Purdue won 15 games}. \end{aligned}$$

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 $\mathbb{P}[A \mid B]$ and $\mathbb{P}[B \mid A]$.

Solution. The following probabilities are easy to calculate:

$$\mathbb{P}[A] = \mathbb{P}[\{\text{⚂}\}] = \frac{1}{6}, \qquad \text{and} \qquad \mathbb{P}[B] = \mathbb{P}[\{\text{⚀}, \text{⚂}, \text{⚅}\}] = \frac{3}{6}.$$

Also, the intersection is

$$\mathbb{P}[A \cap B] = \mathbb{P}[\{\text{⚂}\}] = \frac{1}{6}.$$

Given these values, the conditional probability of A given B can be calculated as

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$$

In other words, if we know that we have an odd number, then the probability of obtaining a 3 has to be computed over $\{\text{⚀}, \text{⚂}, \text{⚅}\}$, 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 {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}, 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

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\frac{2}{12}}{\frac{6}{12}} = \frac{1}{3},$$
$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = \frac{\frac{2}{12}}{\frac{5}{12}} = \frac{2}{5}.$$

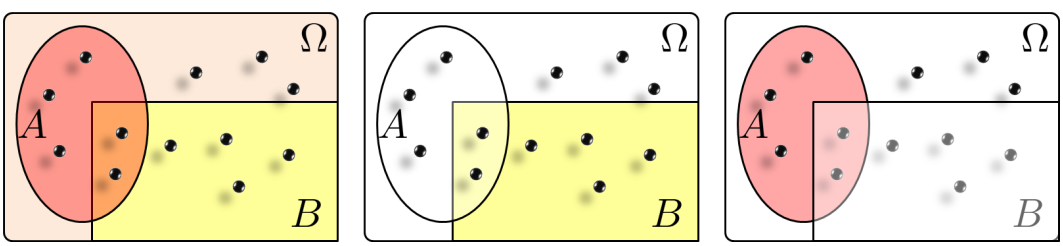


Figure 2.27: Visualization of Example 2.45: [Left] All the sets. [Middle] $P(A|B)$ is the ratio between dots inside the light yellow region over those in yellow, which is $\frac{2}{6}$. [Right] $\mathbb{P}[A|B]$ is the ratio between dots inside the light pink region over those in pink, which is $\frac{2}{5}$.

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

$$\mathbb{P}[M|B] = \frac{\mathbb{P}[M \cap B]}{\mathbb{P}[B]} = \frac{\frac{2}{16}}{\frac{5}{16}} = \frac{2}{5}.$$

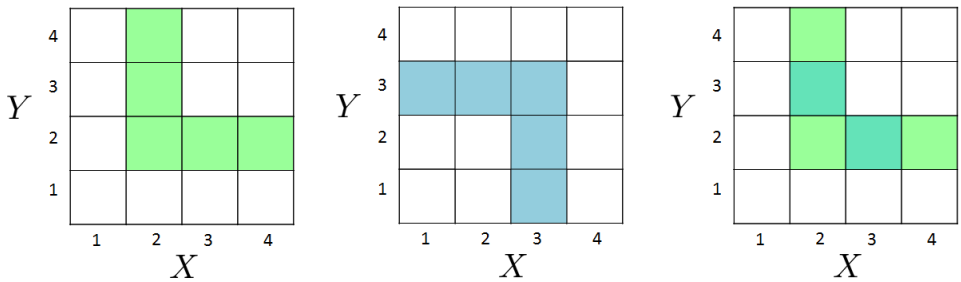


Figure 2.28: Visualization of Example 2.46. [Left] Event B. [Middle] Event M. [Right] $\mathbb{P}(M|B)$ is the ratio of the number of blue squares inside the green region to the total number of green squares, which is $\frac{2}{5}$.

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,

$$\begin{aligned}\mathbb{P}[A \cup C \mid B] &= \frac{\mathbb{P}[(A \cup C) \cap B]}{\mathbb{P}[B]} \\ &= \frac{\mathbb{P}[(A \cap B) \cup (C \cap B)]}{\mathbb{P}[B]} \\ &\overset{(a)}{=} \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} + \frac{\mathbb{P}[C \cap B]}{\mathbb{P}[B]} \\ &= \mathbb{P}[A|B] + \mathbb{P}[C|B],\end{aligned}$$

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 $\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$. If A and B are independent, then $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$ and so

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[A]\,\mathbb{P}[B]}{\mathbb{P}[B]} = \mathbb{P}[A]. \tag{2.29}$$

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*

$$\mathbb{P}[A \mid B] = \mathbb{P}[A] \quad \textit{or} \quad \mathbb{P}[B \mid A] = \mathbb{P}[B]. \tag{2.30}$$

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.

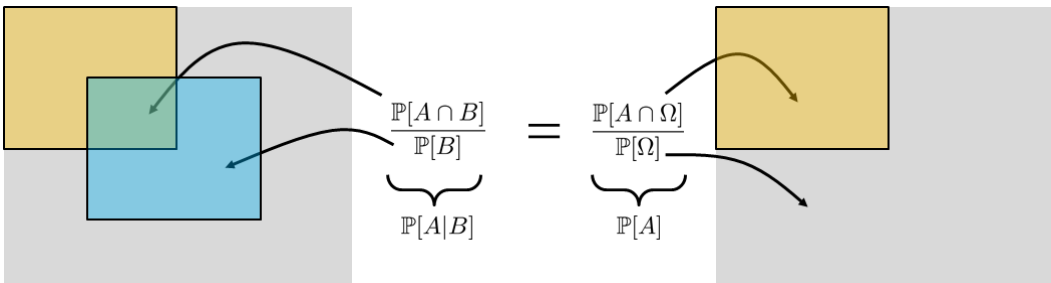


Figure 2.29: Independence means that the conditional probability $\mathbb{P}[A|B]$ is the same as $\mathbb{P}[A]$. This implies that the ratio of $\mathbb{P}[A \cap B]$ over $\mathbb{P}[B]$, and the ratio of $\mathbb{P}[A \cap \Omega]$ over $\mathbb{P}[\Omega]$ are the same.

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.