A minimal introduction to probability theory, statistical inference & hypothesis testing

Jesús Urtasun Elizari

January 29, 2026

Contents

Preface

The purpose of these notes

In the following pages one will find an introductory course to the theory of probability and statistical inference, aiming to cover both foundations and basic mathematical concepts, but also practical tools to deal with real data science problems, such as bayesian probability and hypothesis testing. The text is composed by seven chapters, together with some appendix sections reviewing basic mathematical notions, and a bibliographic note. The purpose of these lecture notes is to make both probability and statistical analysis an easy, engaging and exciting topic for anyone interested, without the need for prior experience.

Both, predictive probability and descriptive statistics have deep historical roots, from ancient works on chance and divination to modern scientific topics oriented towards information theory, modelling and data analysis. As one could guess, rivers of ink have been written about such topics, and endless literature sources are available. However, after following many different courses at both bachelor and postgraduate levels, and teaching such topics myself during the last three years, I have found that most resources belong, almost certainly, to one of the next three classes. Either (*i*) deeply mathematical, and hence out of reach for most experimental or clinically oriented scientists, (*ii*) laboratory oriented, focusing on inference and experimental design, and hence missing most of the mathematical background, or (*iii*) with a direct focus towards programming and computation, relying on domain specific notebooks (Python, R, Matlab, SPSS, etc), and online resources with precompiled libraries for simulation, which again miss most of the mathematical and formal intuitions. Indeed, the misuse of statistics in experimental sciences is a critical topic in modern times, as mathematicians have extensively discussed during the last decades. The well-known article by John P. A. Ioannidis, *"Why most published research findings are false"* [1], serves as a prominent example, and it may serve as motivation for a rigorous study.

As a matter of fact, when it comes to modern statistics, data analysis or experimental design, concepts like *stochasticity, randomness, sampling, hypothesis, significance, statistic test, p-value*—just to mention some of them—are frequently used, but for most bachelor and even master's level degrees they are rarely introduced or properly defined. Indeed, for most experimental and clinically oriented degrees, they are not introduced at all, leaving the student with just a superficial knowledge relying on intuition about some particular cases. Hence, developing high-quality, simple, and accessible open source material for present and future generations, covering both probability and statistical inference from both a fundamental *and* applied level, remains an urgent task for scientists and educators.

This is intended to be a complete introductory course, and no previous mathematical background is required. By keeping the theory simple and always followed by examples, we will build the definitions and quantities from simple to more complex. All mathematical formulas will be introduced with rigorous notation, but keeping in mind that it is not the symbols or the numbers, but the intuitions and the general understanding, what we are after. Additionally, all topics will be introduced alongside with some short historical discussion and context, as we believe that a purely technical knowledge just grasps the complexity—and beauty—of scientific topics. As one

could anticipate already, a proper understanding of ideas such as uncertainty, variation, chance, probability, inference, etc, can be applied to describing a vast amount of real-world phenomena, ranging from gambling and games of chance to data analysis and modelling in physics, biology, machine learning and quantum mechanics, among many others.

As mentioned, the course is organised in five chapters.

- Chapter 1 [...]

- Chapter 2 [...]

- Chapter 3 [...]

- Chapter 4 [...]

- Chapter 5 [...]

- Chapter 6 [...]

- Chapter 7 [...]

At the end of each chapter there will be a series of exercises and coding examples to illustrate and demonstrate the concepts discussed. To avoid misconceptions, let us emphasize here that both, probability and statistics are just branches of mathematics dealing chance and information in random events, *much earlier* than computers, coding languages, Python, R or P-values were even conceived. The data-oriented, practical ways in which probability and statistics are usually taught, relying heavily on computation, is just a consequence of the fact that automatized measurements are nowadays available and trendy in modern times [...].

Example textbooks covering introduction to probability and statistical inference, for further reading:

- A simple, intuitive introduction to statistics with few mathematical concepts is provided in Spiegelhalter's *"The Art of Statistics: How to Learn from Data"* [2].

- A more foundational textbook, with more advanced mathematical approach, can be found at DeGroot and Schervish's *"Probability and Statistics"* [3].

- For a philosophical and historical perspective on probability and statistics, please find Forster and Bandyopadhyay's handbook *"Philosophy of Statistics"* [4].

- A comprehensive introduction with focus on practical applications and modern data analysis tools is can be found at Diez, Barr & Mine *"OpenIntro Statistics"* [5].

- For fundamental concepts in probability and statistics, including random variables, distributions and statistical inference, with practical examples and exercises follow Hossein Pishro-Nik's *"Probability, Statistics & Random Processes"* [6].

Introduction

Even fire obeys the laws of numbers.

— J.B. Joseph Fourier

A bit of history

As one might expect, the origins of probability and related concepts can be traced back to very ancient times. Civilizations such as the Babylonians, Egyptians, and Greeks already encountered uncertainty in various aspects of life, including commerce, games of chance, and divination. Consequently, notions of randomness and stochasticity have deep historical roots. For instance, archaeological findings suggest that the earliest known dice date back over 5,000 years, reflecting humanity's early fascination with chance and unpredictability [7]. Although these cultures had not yet developed a formal mathematical theory of probability, they recognized recurring patterns in random events and attempted to anticipate outcomes through either empirical observation or superstition. For a detailed historical overview, see Florence Nightingale's 1962 manuscript *"Games, Gods and Gambling"* [8].

While classical Greek and Roman philosophers frequently discussed the nature of chance, necessity, and determinism, their inquiries remained primarily philosophical rather than mathematical. From Aristotle to Cicero, many have distinguished between events occurring by chance and those determined by fate, foreshadowing later developments in probability theory [9]. These early ideas, though lacking quantitative formalism, provided the intellectual foundation for later scientific inquiry into randomness and causality.

A significant shift occurred during the late medieval and early Renaissance periods, when more rigorous mathematical ideas began to shape. Italian mathematician and gambler Gerolamo Cardano (1501–1576) made substantial contributions to the mathematical analysis of chance. His work *"Liber de Ludo Aleae (Book on Games of Chance)"* [10], posthumously published in 1663, is one of the earliest known texts to explore probability through the analysis of gambling problems. However, Cardano's reasoning, while insightful, lacked the symbolic clarity and mathematical rigour of modern probability theory. Readers consulting the original manuscript will notice an ambiguous and sometimes inconsistent symbolic system, quite unlike the formal structures we use nowadays.

The formalization of probability as a mathematical discipline did not occur until the 17th century, most notably through the seminal correspondence between Blaise Pascal and Pierre de Fermat. Their work, motivated by problems such as finding a fair division of stakes in interrupted games of chance, introduced foundational concepts such as combinatorics, expected values, and variance [11]. These developments paved the way for later contributions by Christiaan Huygens, who in 1657 wrote the first published textbook on probability *"De Ratiociniis in Ludo Aleae (On Reasoning in Games of Chance)"* [12], and Jacob Bernoulli, whose 1713 *"Ars Conjectandi (The Art of Conjecturing)"* remains among the most influential early texts in the field. Their works, along with many others, collectively laid the groundwork for the probabilistic and statistical

methods that foreshadow modern scientific reasoning [13, 14].

From the 19th century onwards, probability theory became increasingly intertwined with statistics and inference, giving rise to the mathematical and empirical frameworks that led to modern scientific analysis of uncertainty. Building on the analytical foundations laid by Laplace—who unified probability with inference through inverse probability—and Gauss's theory of errors, Adolphe Quetelet—Belgian astronomer and mathematician—played a decisive role in extending probabilistic reasoning to social and biological phenomena. His notion of the statistical individual, or *homme moyen*, framed variation not as noise to be eliminated, but as a fundamental object of study, helping to establish statistics as a distinct discipline concerned with populations rather than isolated events [15, 16, 17].

Within this emerging landscape, Florence Nightingale stands as a central figure in the practical and institutional adoption of statistical reasoning. Through her innovative use of graphical representations and her advocacy for quantitative evidence in public health policy, she demonstrated the power of statistical methods as tools for decision-making under uncertainty [18]. In parallel, Joseph Fourier's work on heat conduction introduced series expansions and integral transforms that, while originally developed in the context of mathematical physics, would later become indispensable in the study of random processes, signal analysis, noise, and diffusion. Although Nightingale and Fourier approached uncertainty from fundamentally different directions—one through empirical data and social reform, the other through mathematical analysis of physical systems—their contributions expanded the scope of probabilistic thinking and helped prepare the intellectual ground for the later development of stochastic processes and statistical physics.

A further conceptual leap occurred in the early 20th century with the work of Andrey Markov, a Russian mathematician working at a time when probability theory was still largely built on the assumption of independent events. Markov was interested in understanding whether the familiar regular patterns of probability—such as averages stabilizing over time—could still arise when events influenced one another. To explore this question, he studied sequences in which the outcome of each step depends only on the immediately preceding one, rather than on the entire past.

This simple idea, later formalized as what are now called Markov chains, showed that randomness and order need not rely on complete independence. Even when successive events are linked, long-term statistical patterns can still emerge under suitable conditions. Markov's work, developed within the strong Russian tradition of rigorous analysis, opened the door to the modern study of systems that evolve over time under uncertainty, from physical processes to language, biology, and many contemporary data-driven models [19].

Markov's investigations inaugurated the systematic study of dependence in stochastic processes and laid the groundwork for much of modern probability theory. By showing that long-run behavior could be analyzed without recourse to independence, his work opened new avenues in the mathematical treatment of dynamical systems subject to randomness. The conceptual framework he introduced has since become central to a wide range of disciplines, including statistical mechanics, linguistics, finance, and more recently, machine learning and data science, where Markovian models serve as fundamental tools for modeling sequential and temporal data [20].

As a final note, the modern axiomatic formulation of probability was introduced in the early 20th century by the Russian mathematician Andrey Kolmogorov. In his 1933 monograph *"Grundbegriffe der Wahrscheinlichkeitsrechnung (Foundations of the Theory of Probability)"* [21], Kolmogorov synthesized classical and frequentist ideas into a rigorous mathematical framework based on measure theory. His axioms remain the standard foundation for probability theory to this day. It may seem surprising that a concept with such ancient origins was not formally axiomatized until relatively recent times, and we will return to Kolmogorov's formulation

and its implications in greater detail in further chapters. Nevertheless, philosophical discussions about the interpretation of probability and its relation to the physical sciences—especially in the context of determinism, epistemology, and modern topics such as quantum mechanics—predate Kolmogorov's formulation and continue to evolve to this day.

Chapter 1

Descriptive statistics

Statistics is the grammar of science.

— Karl Pearson

As a first approach to probability and statistics, we should properly define both topics and their main fields of study. Although deeply related, and both historically rooted in *combinatorics*—the mathematical study of counting discrete structures, and how they change—they constitute well-differentiated fields of mathematical analysis. A clear distinction often made is that probability is a *predictive* branch of mathematics, dealing with random events, and aiming to compute expected values for unknown outcomes. On the other hand, statistics can be viewed as a *descriptive* approach to uncertainty, based on sampling finite sets of observations from a given population and constructing informative quantities, called *estimators*, to explore central tendency and variation. Such distinctions have been extensively debated by mathematicians, experimental scientists, and philosophers of science [22].

As a rule of thumb, probability provides a formal language for modelling uncertainty, whereas statistics concerns the epistemic problem of learning from data. In this chapter, we introduce basic ideas of statistical inference such as population, sampling, and estimators of central tendency and variation, together with notions of representation and visualization. The foundations of probability theory, rooted in the works of Bernoulli, Laplace, and Gauss, among others, will be covered in Chapter 2.

A philosophical position often adopted is that statistics is essentially the study of uncertainty, and that the statistician's role is to assist other fields that encounter uncertainty in their work. In practice, statistics is ordinarily associated with data, and it is the link between the variability in the data and the uncertainty inherent in the phenomenon under study that has occupied statisticians [23].

As a final note, let us emphasize how these two approaches can and do coexist in science. It is often said that science proceeds by formulating hypotheses and making predictions, which are then compared and benchmarked against experimental results. While this description applies well to many areas, it is a simplification and not universally accurate. Some sciences—such as Newtonian mechanics, much of physics, and chemistry—rely on building precise models that generate quantitative predictions later tested by experiment. A clear example is the use of Newtonian mechanics to predict where and when a stone will fall when thrown, followed by experimental measurement.

By contrast, a paradigmatic example of an inference-driven scientific theory is Darwinian evolution, which does not aim primarily to predict individual outcomes but rather to reconstruct and explain observed patterns from available evidence. This distinction is worth emphasizing, as definitions of science that focus exclusively on predictive power can be misleading. Different sciences may differ substantially in methods, instrumentation, and conceptual tools, yet they

are all equally legitimate in their aims and explanatory roles [24, 25].

1.1   Population and sampling

A large part of the history of science can be seen as a continuous effort to translate observations of reality into precise mathematical terms. Describing natural phenomena in numerical language requires tools that relate one or more relevant quantities—often called *variables*—and explain how they change with respect to one another. The goal of modelling may be, for example, to determine the distance from the Earth to the Sun, estimate the number of stars in the observable universe, or relate the incidence of lung cancer to environmental or behavioral factors, such as smoking.

prediction (probability)

theory   experiment
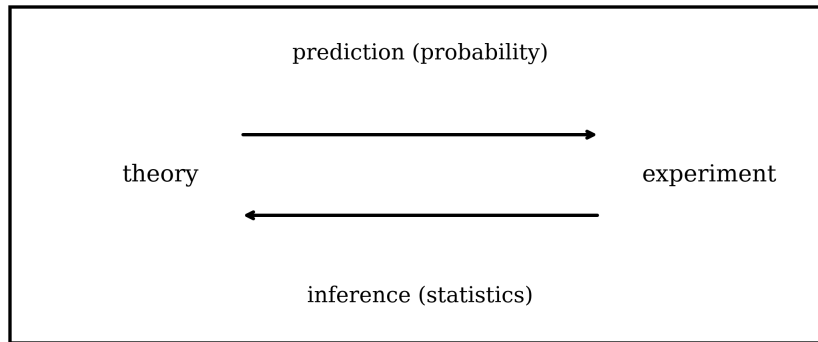
inference (statistics)

Figure 1.1: Representation of two complementary approaches to randomness and variability: a predictive, theory-driven path from model to experiment, and a descriptive, inference-based path from data to explanation. Probability and statistics are normally framed as the branches of mathematics addressing both ways.

In the same way mathematics is often summarized as the tasks related to *count*, *measure*, and *sort*, statistical problems can be grouped into three broad categories. *Sampling* concerns the selection of a finite set of observations from a larger, typically unknown population. *Estimation* involves constructing numerical summaries that describe how the data are distributed, while *visualization* addresses how observations are represented and how such representations influence interpretation. All three are fundamentally concerned with uncertainty and variability.

Hence, all statistical inquiries begin with observations and measurements, which we normally refer to as *data*. And data begins with the act of selection, or *sampling*. The natural world offers a vast abundance of phenomena, with endless opportunities for observation, but only a finite subset can ever be recorded. This distinction gives rise to two central notions: the *population*, which we denote by $\mathcal{P}$, represents the complete set of all possible observations under study. We will write it as

$$\mathcal{P} = \{x_1, x_2, \ldots, x_N\} \,. \tag{1.1}$$

The *sample* $\mathcal{S}$, on the other hand, is the finite subset actually collected. For a series of N observations x_1, x_2, ..., x_N, a sample of just n elements—less than the total, which is denoted by the upper case N—is defined as

$$\mathcal{S} = \{x_1, x_2, \ldots, x_n\}, \quad n < N \,, \tag{1.2}$$

The population represents the ideal object of inference, while the sample is the concrete, finite evidence available to us. As an example, if I want to study some disease and its relation smokers in a given country, the behavior of a specific population, or the active genes in the genome, I will never have access to the *complete population*, but only to the amount of them that I am able to

question, measure, or survey. This distinction is far from trivial. A poorly chosen sample often misrepresents the population and may induce bias, whereas a carefully constructed one mirrors its essential features, and can be used to describe the underlying nature.

Equally important is the recognition that not all data is equal, neither behaves in the same way. A common distinction is to consider *categorical* and *numerical* data. Categorical—or *qualitative*—data describes qualities or labels such as the eye colour of students in a classroom (blue, brown, green), the brand of a purchased smartphone, etc. Sometimes they are further divided into *nominal* categories, with no natural order, like the eye colour or the smartphone brand, and *ordinal* categories with a meaningful order. Examples of these would be the finishing places in a race (first, second, third), survey responses ranging from *strongly disagree* to *strongly agree*, etc.

The other big family is normally referred to as numerical—or *quantitative*—data. It represents numerical quantities and is often subdivided into *discrete*, countable numbers, such as the number of books on a shelf (4, 5, 6) or the number of goals scored in a match, and *continuous* values that can take any number within a range, such as the time a sprinter takes to run 100 meters, or the height of a person measured with some arbitrary precision.

Distinguishing between these types is no mere slang; different types of observations require different mathematical tools, and will be described in different ways. For example, it would not make sense to compute a mean out of smartphone brands, but to compute the mean of their prices is informative. Similarly, the distribution of finishing places after a race might be summarized by a median position, whereas heights of athletes could be studied with averages and measures of spread. A correct classification of data is thus a safeguard against misuse and a guide toward insight.

As a summary, sampling and proper description of data establish the ground upon which statistics is built. Before calculating, summarizing, or diving into inference, one must ensure that the information collected is both *representative and properly understood*. Without these foundations, descriptive measures risk floating unmoored, detached from the reality they claim to represent. Accurate sampling and rigorous description will lead to a faithful representation of the phenomena under study and their relationships, detecting anomalies, and even building accurate predictions.

Let's end this section with a historical note. As we have mentioned, uncertainty has long been associated with games of chance and gambling, but it was not addressed as a statistical problem until much later. The Royal Statistical Society, founded in 1834, together with many other early statistical organizations, was originally established to gather and publish data, with the aim of informing social and economic questions through systematic measurement. It did not take long before statisticians began to ask how such data might best be analyzed and interpreted, and modern *statistical inference* gradually emerged. Among the Society's influential early figures were Adolphe Quetelet, whose work helped establish statistics as a tool for studying social phenomena, and Charles Babbage, who advocated quantitative approaches to scientific and administrative problems.

Among its most famous members was Florence Nightingale, admitted in 1858 as the Society's first female member, whose work exemplified the practical use of statistical reasoning in policy and public health. In the early 20$^{\text{th}}$ century, foundational questions concerning probability and inference were further clarified through the work of thinkers such as Frank P. Ramsey, who connected probability with rational belief and decision-making [26]. Other notable presidents of the Royal Statistical Society have included William Beveridge and Ronald Fisher, whose contributions will be discussed in Chapter 4.

Andrew Lang's famous quote *"most people use statistics as a drunken man uses lamp-posts—for support rather than illumination"*, highlights the tendency to use statistics as a crutch,

relying on them for validation rather than seeking genuine understanding. Lang's observation serves as a cautionary reminder to approach statistical data with critical thinking and not merely as a tool to bolster preconceived notions.

1.2   Central tendency and variation

Once we have drawn a clear distinction between the population under study and the selected sample, we immediately face a fundamental problem. Neither the population mean—often referred to as the *true* mean and usually denoted by μ—nor the population variance, known as the *true* variance and written as σ^2, are directly available to us. As we have seen, the only information at our disposal is the finite set of observations contained in our sample. From this limited information, we therefore seek to construct quantities that provide insight into key features of the population, such as its central value or its variability. These quantities are known as *statistical estimators*. Common examples include the *sample mean*, the *median*, the *variance* and the *standard deviation*.
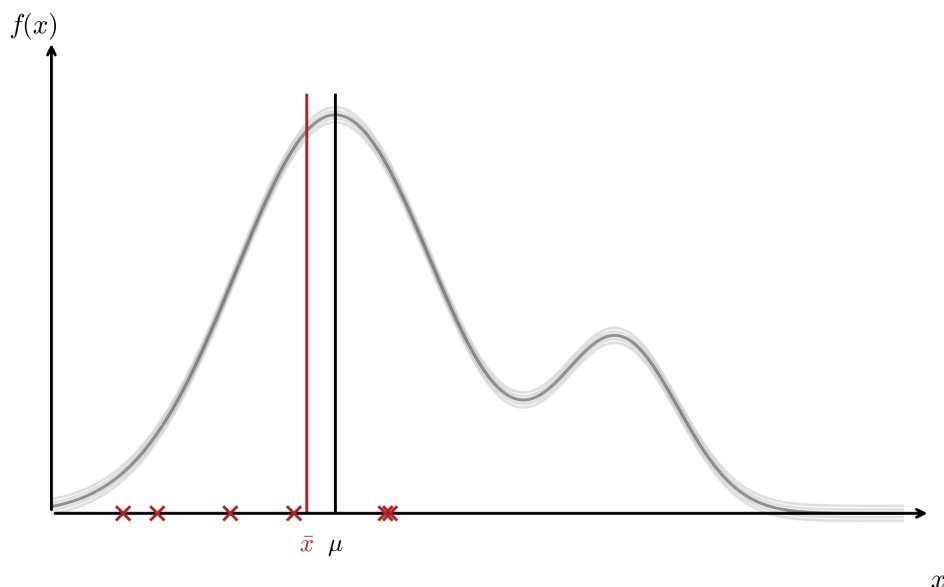

Figure 1.2: Representation of the *true* population mean μ, in black, and the observed *sample* mean $\bar{x}$. The true mean is and ideal and unaccessible quantity, while the sample mean can be computed as an estimator of the finite sample.

Once observations have been collected, a natural question arises: what is the *center*, or *typical*, value of this data set? Measures of central tendency address this question by summarizing the data with a single representative number, offering an immediate sense of where the observations are located within the distribution.

The sample mean:

The *sample mean*, or *average* is perhaps the most familiar measure of central tendency. Imagine we are doing an experiment where we measure some variable, and let's call it x for simplicity. x can be anything we could measure, like position at a given time, energy of some system, concentration of a specific substance, etc. Let's imagine we repeat the measurement n times, and we obtain the values $x_1, x_2, \ldots, x_n$. That will be our set of observations—our *sample*—$\mathcal{S}$. We could simply write it as a list—or a *vector*—in the following way:

$$\mathcal{S} = \{x_1, x_2, \ldots, x_n\} \, .$$

Keep in mind that from the mathematics perspective the word *vector* has a slightly different meaning, with subtleties related to algebraic operations and relations they should satisfy, but for the purpose of this course, where we prioritize above all simplicity, a vector and a list of numbers will be essentially the same thing.

We can define the sample mean of an arbitrary large sample of n observations, as the sum of all elements divided by the total. We will write it as $\bar{x}$, and define it as follows:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) \, . \tag{1.3}$$

We can write this in a slightly more compact way as a *summation*, as follows:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \, . \tag{1.4}$$

Here we denote the sum of all elements x_i with the greek letter $\sum$, starting with the first one (x_1, for $i = 1$) and until the last one (x_n, for $i = n$). The expressions (1.3) and (1.4) mean *exactly* the same thing, just written in different ways.

Let's illustrate with an example. Suppose we repeat a measurement three times, obtaining the results $x_1 = 1$, $x_2 = 2$, and $x_3 = 3$. Our sample is then $\mathcal{S} = \{1, 2, 3\}$, and the sample mean is

$$\bar{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{1}{3}(1 + 2 + 3) = 2 \, .$$

As a warm-up exercise, try computing the same mean value for a second sample, let's say $\mathcal{S} = \{4, 5, 6\}$. Substituting into the general expression (1.4) gives

$$\bar{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{1}{3}(4 + 5 + 6) = 5 \, .$$

As we see, the sample mean captures information about the "central" value, where most events cluster. Although useful, it is sensitive to extreme values—often called *outliers*—which motivates the definition of additional, more robust measures of central tendency.

The median:

The *median* represents similar information, as the value that splits the ordered data set in half. For an ordered sample $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, the median M is defined as

$$M = \begin{cases} x_{(k+1)} \, , & \text{if } n = 2k + 1 \text{ (odd)} \, , \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} \, , & \text{if } n = 2k \text{ (even)} \, . \end{cases} \tag{1.5}$$

Note that here k is just an integer that helps locate the middle position of an ordered data set of size n. If the sample size n is even, we write $n = 2k$, while for n odd, we write $n = 2k + 1$. In the case of an odd-sized sample, the median is just the middle-point, while for an even size, it is computed as the average of the two middle points.

The mathematical definition (1.5) may seem a bit unnatural at first, so let's navigate it with a couple of examples. Consider the sample $\mathcal{S} = \{1, 2, 3, 5, 3, 2, 7\}$. First, we order the data:

$$\mathcal{S}_{\text{ordered}} = \{1, 2, 2, 3, 3, 5, 7\} \, .$$

Since the sample has an odd number of elements ($n = 7$), the median is just the middle value:

$$M = x_{(4)} = 3 \, .$$

Now consider an even-sized sample $\mathcal{S} = \{1, 2, 3, 5, 4, 3, 2, 7\}$. Ordering the data gives

$$\mathcal{S}_{\text{ordered}} = \{1, 2, 2, 3, 3, 4, 5, 7\}.$$

Which has now an even number of elements ($n = 8$). Hence, applying such case in (1.5), the median is the average of the two middle values

$$M = \frac{x_{(4)} + x_{(5)}}{2} = \frac{3 + 3}{2} = 3 \ .$$

Unlike the mean, the median is robust to outliers and skewed data, capturing the central position of the dataset even with repeated values. To illustrate that, let's have a look at the following sample $\mathcal{S} = \{1, 2, 3, 3, 4, 4, 200\}$, which contains the value 200 as a huge outlier. The sample mean would be

$$\bar{x} = \frac{1}{7}(1 + 2 + 3 + 3 + 4 + 4 + 200) = \frac{217}{7} = 31 \ .$$

While the median, given a size $n = 7$ would just be the middle (4th) value

$$M = 3 \ .$$

For instance, the data represented in LHS of Figure 1.4 will be accurately described by computing the mean, given its symmetric behaviour, while the one in the RHS will be better addressed with a median, accounting for the skewness and the presence of outliers.

The sample variance:

Beyond central location, it is important to understand the *spread* of the data. We define the *sample variance* s^2 of a data set as a quantity that captures how far the elements are from the mean value:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \ . \tag{1.6}$$

The factor $n-1$ in the denominator of (1.6) is called the *Bessel correction factor*. A technical explanation is that this correction ensures that s^2 is an *unbiased estimator* of the population variance, a concept we will discuss in Chapter 3. In particular, it reflects the fact that the sample mean $\bar{x}$ is itself estimated from the data.

Note that the variance is obtained by summing the squared differences between each observation and the mean. The differences are squared so that positive and negative deviations do not cancel and the result is always non-negative. If all elements in the sample are very close to the mean, the variance s^2 will be small; if the elements are widely spread, the variance will be larger.

Let us illustrate this definition with an example. For the sample $\mathcal{S} = \{1, 2, 3\}$, which has $n = 3$ observations and sample mean $\bar{x} = 2$, the variance is

$$s^2 = \frac{1}{3-1} \sum_{i=1}^{3} (x_i - \bar{x})^2 = \frac{1}{2}\big((1-2)^2 + (2-2)^2 + (3-2)^2\big) = \frac{1}{2}(1 + 0 + 1) = 1 \ ,$$

which we could interpret as, on average, the elements of the list being *one unit* away from the mean.

As a warm-up exercise, try to compute the variance for a second sample, let's say $\mathcal{S} = \{4, 5, 6\}$. By substituting in the general expression (1.6) you should get the result

$$s^2 = \frac{1}{3-1} \sum_{i=1}^{3} (x_i - \bar{x})^2 = \frac{1}{2}\big((4-5)^2 + (5-5)^2 + (6-5)^2\big) = \frac{1}{2}(1 + 0 + 1) = 1 \ .$$

We obtain again a variance $s^2 = 1$, indicating as in the previous example, that the elements of this sample $\mathcal{S}$ are also *one unit* away from the mean.

The standard deviation:

Another useful quantity used to characterize variability is the so-called *standard deviation*, which is simply the square root of the variance,

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \, , \tag{1.7}$$

and therefore has the same units as the original data. While the variance measures spread in squared units, the standard deviation provides a more interpretable measure of how far, on average, observations lie from the mean.

At a glance, variance and standard deviation both quantify how much the elements of a data set deviate from the mean, capturing the notion of *spread*.

Quantiles:

Finally, *quantiles* provide another way to describe the distribution of data by dividing an ordered data set into equal proportions. The p-th quantile Q_p is defined as the value below which a fraction p of the data lies. Common special cases include the *first quartile* (Q_1, corresponding to the 25th percentile), the *median* (Q_2, the 50th percentile), and the *third quartile* (Q_3, the 75th percentile). Informally, they answer questions of the form: *"What value separates the lowest $p\%$ of observations from the rest?"* They are particularly useful for describing skewed data and for comparing distributions without relying solely on the mean. To illustrate this idea with a simple example, consider the ordered sample

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7, 8\} \, .$$

Since there are $n = 8$ observations, the median Q_2 lies halfway between the fourth and fifth elements and is therefore

$$Q_2 = \frac{4+5}{2} = 4.5 \, .$$

The first quartile Q_1 separates the lowest 25% of the data and lies between the second and third elements, while the third quartile Q_3 lies between the sixth and seventh elements. In this way, quantiles summarize the distribution by marking its lower tail, central region, and upper tail.

A rigorous definition of quantiles requires the notion of a distribution function and cumulative probability, which we will introduce in the next chapter. For a continuous cumulative distribution function (CDF) F, the p-th quantile satisfies

$$Q_p = \inf\{x : F(x) \geq p\} \, . \tag{1.8}$$

In summary, the mean, median, mode, variance, standard deviation, and quantiles provide a rich and complementary view of a data set's central tendency and variability. Together, they allow for both numerical and graphical summaries that capture essential features of the data.

Variation is not merely a technical detail; it is the essence of uncertainty. Without spread, probability would be trivial, as every outcome would be identical. It is precisely in the differences among observations that statistical inquiry finds its substance. Central tendency and variation thus form complementary lenses through which data become intelligible: they allow us to assess whether groups are similar or different, whether an observation is ordinary or surprising, and whether observed variation can plausibly be attributed to chance. In this sense, descriptive statistics foreshadows the inferential methods to come, hinting at deeper regularities beneath the numbers.

1.3 Data visualization

While numerical summaries are useful, the human mind often grasps patterns much more quickly through vision than through calculation. By *data visualization* we mean a family of techniques used to transform numbers and sequences into shapes, colours, and spatial structures that are easier to interpret and can often be understood at a glance. Visualization turns abstraction into perception and frequently reveals regularities that remain hidden when data are examined only through formulas or numerical summaries. Today, a broad range of disciplines grouped under the name of data visualization—or data *representation*—have become central pillars of scientific practice and data-driven inquiry.

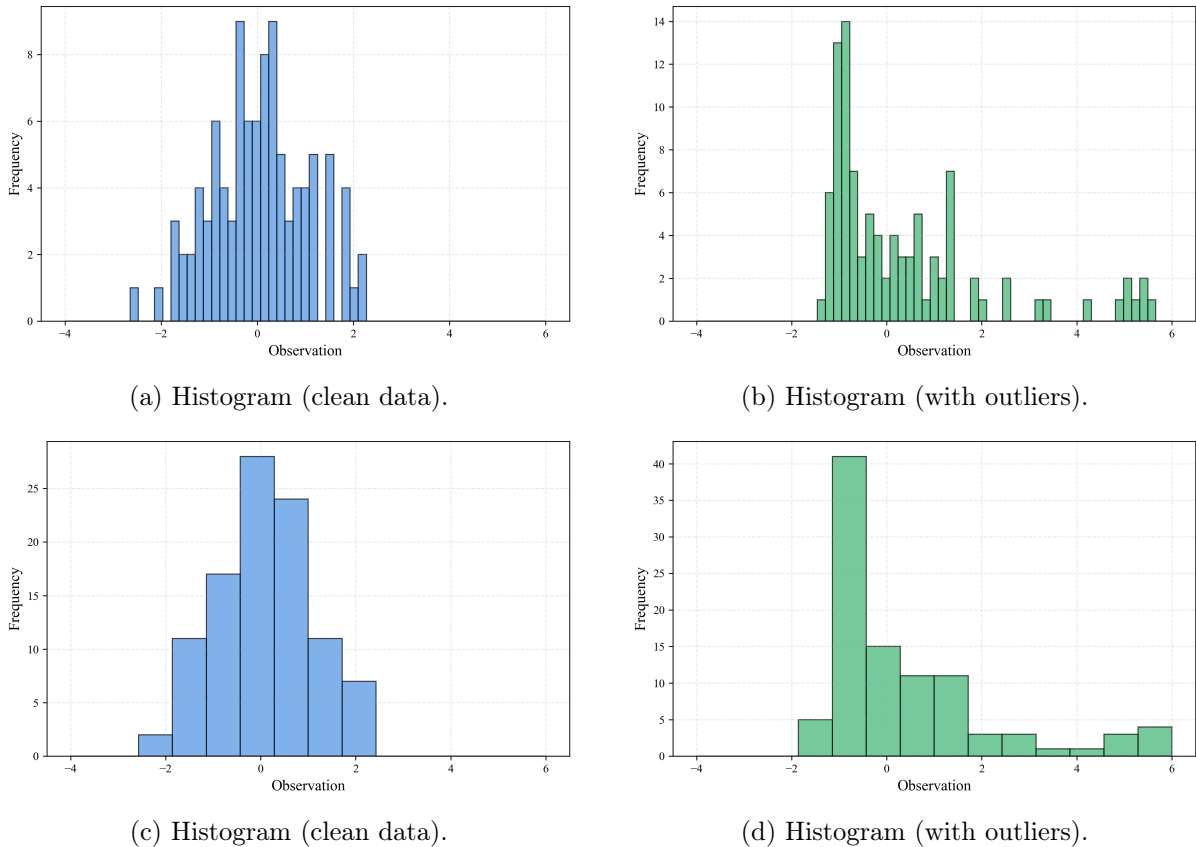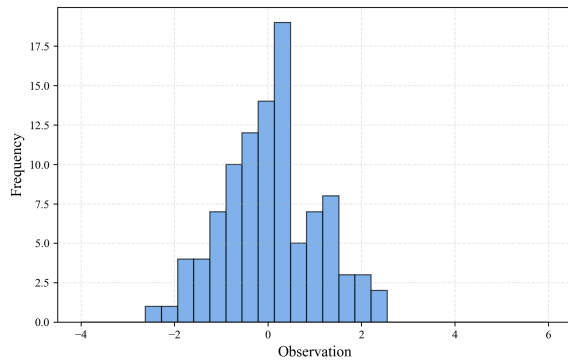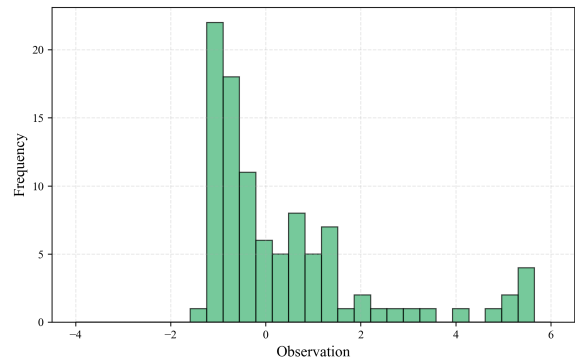| (a) Histogram (clean data).	(b) Histogram (with outliers).
(c) Histogram (clean data).	(d) Histogram (with outliers).

Figure 1.3: Comparison of graphical summaries for $n = 100$ observations drawn from a Gaussian distribution. Left column: uncontaminated data, with both small and large binning size. Right column: data affected by outliers, with both small and large binning size. Figure shows how a small bin size may give lead to resolution, but also introduce noise.

Among the oldest and most fundamental visualization tools is the *histogram*. The idea of dividing data into intervals in order to visualize frequency dates back to the late 19th century, when Karl Pearson formalized the histogram as a graphical representation closely connected to probability distributions [27, 28]. A histogram divides the range of a data set into consecutive intervals, or *bins*, and represents the number—or relative *frequency*—of observations falling within each bin by the height of a bar. This simple yet powerful plot provides an immediate visual impression of the distribution, allowing one to identify symmetry, skewness, concentration of values, and potential gaps. For instance, a symmetric histogram suggests a roughly balanced distribution around the mean, whereas a right-skewed histogram indicates that large values are less frequent but may still exert a strong influence on measures such as the mean. Histograms are therefore widely used in the physical sciences and in contexts where data are compared to
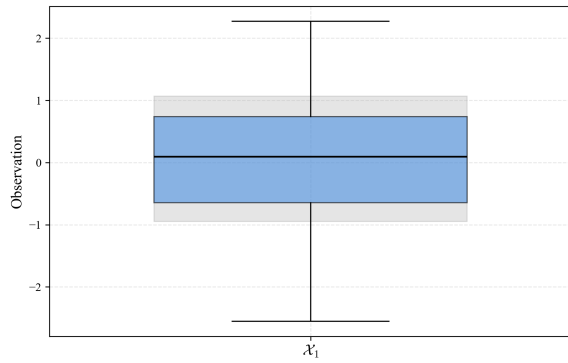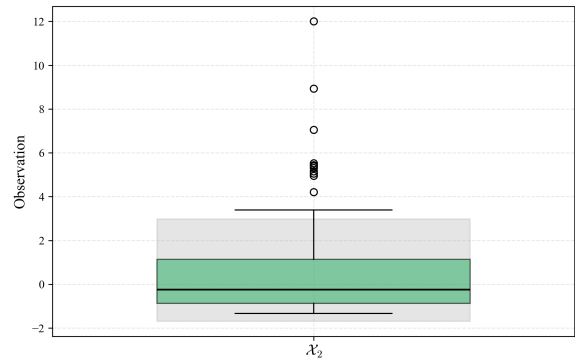
theoretical or mathematical predictions.
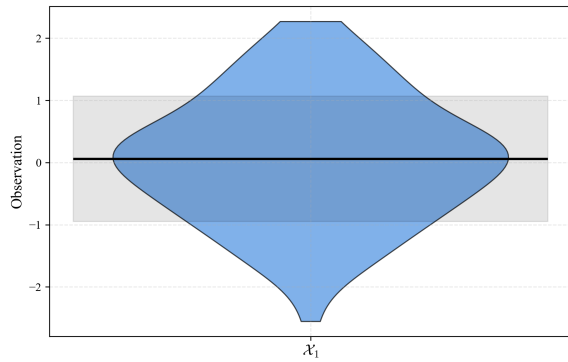

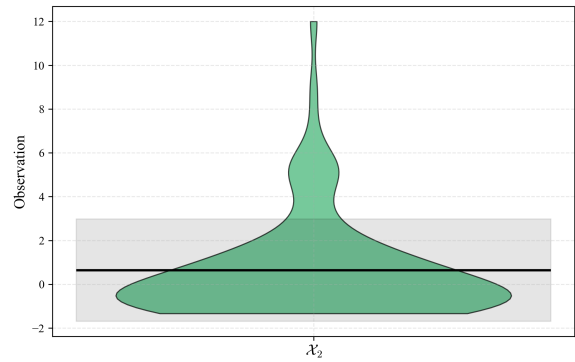(a) Histogram (clean data).

(b) Histogram (with outliers).

(c) Box plot (clean data).

(d) Box plot (with outliers).

(e) Violin plot (clean data).

(f) Violin plot (with outliers).

Figure 1.4: Comparison of graphical summaries for $n = 100$ observations drawn from a Gaussian distribution. Left column: uncontaminated data. Right column: data affected by outliers. The rows show, respectively, histograms, box plots, and violin plots, illustrating how different visualization techniques respond to skewness and extreme observations.

Constructing an informative histogram requires some care. As a general rule, bins should be of equal width and together cover the entire range of the data, while reflecting natural groupings when possible. The choice of bin width plays a crucial role: using many narrow bins increases resolution but can introduce spurious fluctuations, especially in the presence of outliers, whereas using a small number of wide bins produces a smoother and more robust picture at the cost of detail. Because extreme values can strongly affect the appearance of a histogram, they must be handled thoughtfully. For highly skewed distributions, numerical summaries such as the median and the interquartile range (IQR) often provide a more stable description than the mean.

Another widely used visualization is the *box plot*, also known as the *box-and-whisker* plot, introduced by John Tukey in 1970 as part of his work on exploratory data analysis [29]. The box plot offers a compact summary of a data set's central tendency, spread, and potential outliers. It is constructed from five key statistics: the minimum, first quartile (Q_1), median (Q_2), third quartile (Q_3), and maximum. The box itself represents the interquartile range (IQR = $Q_3 - Q_1$), while observations lying more than 1.5 times the IQR from the quartiles are typically flagged as outliers. Box plots allow for rapid comparisons across multiple groups and are particularly effective at revealing asymmetry and variability without being overly influenced by extreme values. For this reason, they are especially common in biological, medical, and clinical sciences, where experiments are often repeated many times with relatively small sample sizes.

More recently, the *violin plot* has emerged as a refinement of the box plot, combining summary statistics with a smooth representation of the distribution's shape. Although its precise origin is less sharply documented, the violin plot gained prominence in the late 20th century with the development of statistical software environments such as R and Python, particularly following the work of Hintze and Nelson in the 1990s [30]. A violin plot augments the box plot by incorporating a kernel density estimate of the data, producing a mirrored shape that reflects the underlying distribution. In addition to displaying the median and quartiles, violin plots reveal features such as skewness and multimodality that may be obscured in simpler summaries, making them especially useful for comparing several groups side by side.

The importance of visualization extends beyond convenience. Early work by Spearman on association [31] and later demonstrations such as Anscombe's quartet [32] highlighted a crucial lesson: data sets with nearly identical numerical summaries can exhibit dramatically different graphical structures. Visualization is therefore not merely a tool for presentation, but a fundamental instrument of statistical reasoning. By complementing numerical summaries, graphical representations help guard against misinterpretation and deepen our understanding of variability, structure, and uncertainty in data.

As a note, let us emphasize that the roots of data visualization can be traced back well before modern statistics, to the early development of analytical geometry in the 17th century. A decisive step was taken by René Descartes, whose introduction of coordinate systems in *"La Géométrie"* (1637) provided a systematic way to represent numerical relationships geometrically [33]. By associating quantities with positions in space, Descartes established the conceptual framework underlying graphs, curves, and plots as representations of functional relationships. Although his work was not statistical in intent, it laid the mathematical foundation upon which later visual representations of empirical data would be built. Over the following centuries, this geometric perspective was gradually adapted to observational and social data, culminating in the 19th-century development of statistical graphics—such as frequency plots and histograms—as tools for exploring variability, association, and uncertainty.

Following Descartes' geometric framework, graphical representations of data evolved through the work of figures such as William Playfair, who in the late 18th century introduced line and bar charts to depict economic time series, and Adolphe Quetelet, who applied statistical regularities to social phenomena. In the 19th century, visualization became both an analytical and persuasive tool, exemplified by Florence Nightingale's use of statistical graphics in public health and Charles Joseph Minard's multivariate maps. These developments established visualization as a central instrument for exploring, communicating, and reasoning about data, long before the formalization of modern statistical theory [34, 35].
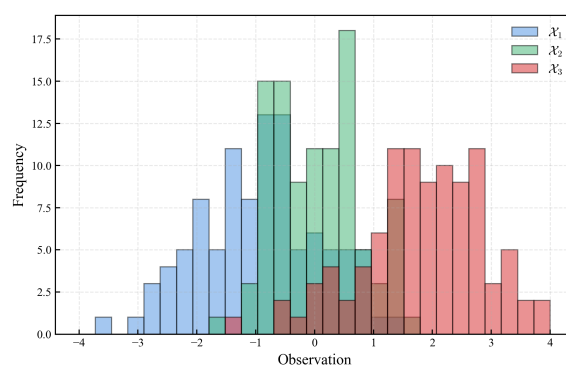
(a) Three sets of observations, with the mean value and standard deviation represented as a histogram-based plot [...].
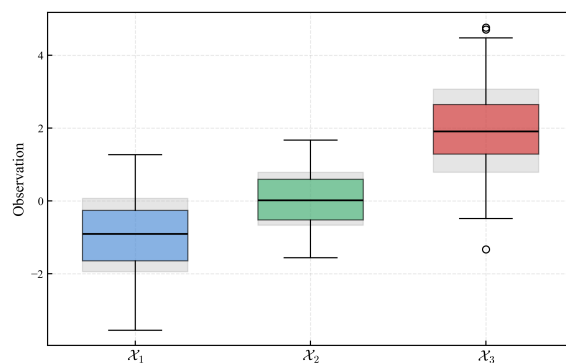
(b) Three sets of observations, with the mean value and standard deviation represented as a box plot [...].
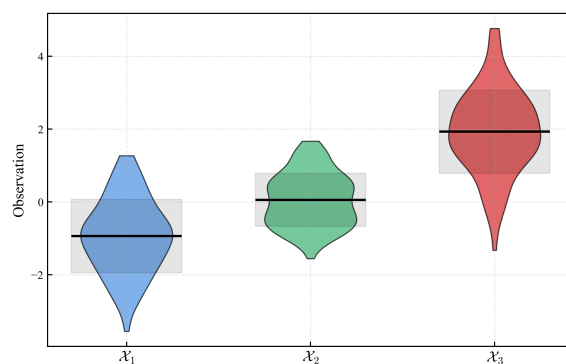
(c) Three sets of observations, with the mean value and standard deviation represented as a violin plot [...].

Figure 1.5: Comparison of three visualization methods—histogram, box plot, and violin plot—showing the mean and variability of three samples of size $n = 100$.

Chapter 2

Foundations of Probability

> *It is through the calculation of probabilities that the divine order becomes visible.*
>
> — Jacob Bernoulli

The study of probability, though having very ancient roots, began its modern development in the 17th century through the famous correspondence between Blaise Pascal and Pierre de Fermat. Their discussion of games of chance, and in particular the "problem of the division of stakes", laid the groundwork for a systematic mathematical analysis of uncertain events. A few decades later, Jacob Bernoulli's *"Ars Conjectandi"* provided the first sustained theoretical treatment of probability, including an early formulation of the law of large numbers [13, 14]. Subsequent refinements by Abraham de Moivre, particularly his work on normal approximation, and by Pierre-Simon Laplace transformed probability into a powerful analytical theory, while its fully axiomatic structure only crystallised in the 20th century, as we will see [16].

Beyond games of chance, probability rapidly became essential for the understanding of natural phenomena and human affairs. Astronomy, population studies, and various physical problems required tools to reason quantitatively about variability, error, and incomplete information. During the 19th century, thinkers such as Cournot emphasized the connection between probabilistic laws and empirical regularities, arguing that probability acquires meaning through its relation to observable frequencies in the world. In this sense, probability emerged not merely as a mathematical curiosity, but as a response to practical problems involving uncertainty and regularity in the empirical world.

As we have mentioned already, a decisive step toward mathematical rigor was taken by Andrey Kolmogorov in 1933. In his *"Grundbegriffe der Wahrscheinlichkeitsrechnung"* [21], Kolmogorov showed that probability could be treated as a branch of measure theory, independent of any specific interpretation. This formulation built directly on earlier developments in analysis, most notably Henri Lebesgue's theory of integration, which provided the mathematical language needed to define probabilities as measures on sets [36]. Rather than defining probability through intuition, symmetry, or frequency, Kolmogorov postulated a small set of axioms from which the entire formal theory follows.

The measure-theoretic approach introduced by Kolmogorov was later systematized and widely disseminated through modern mathematical treatments of probability, most famously in Patrick Billingsley's *Probability and Measure* [37]. These works established the now-standard framework in which random variables, expectations, and convergence are treated using the tools of modern analysis.

Two influential strands in the interpretation of probability emerged from these debates. On the one hand, the frequentist program, most clearly articulated by Richard von Mises,

grounded probability in long-run relative frequencies of repeatable experiments. On the other hand, approaches originating with Thomas Bayes and later developed in the twentieth century interpreted probability as a rational degree of belief. Both perspectives play an important role in modern probability and statistics, and their mathematical and philosophical foundations will be examined in detail in Chapter 5.

Parallel developments in the early 20$^{\text{th}}$ century focused on statistical inference from data rather than individual decision-making. These approaches emphasized long-run frequency properties, error control, and sampling distributions, leading to the classical framework of statistical inference that still underlies much of modern applied statistics.

2.1 Probability and random events

At its heart, probability is nothing more—and nothing less—than a branch of mathematics developed to describe random phenomena, also referred to as *stochastic*. The word "stochastic" comes from the Greek στοχαστικός, literally meaning "to guess" or "to aim." Probability thus provides a numerical language for uncertainty, allowing us to quantify how surprising or plausible an outcome is before it is observed.

In modern mathematics, probability is defined axiomatically following Kolmogorov's axioms [21]. Without much technicality, probability $\mathbb{P}$ is a number we associate to each event, satisfying three fundamental rules:

- Probabilities are never negative: $\mathbb{P}(A) \geq 0$ for any event A.

- The probability of a certain event is 1.

- If two events cannot occur together, the probability that one or the other occurs is the sum of their probabilities.

For a discrete set of all possible outcomes $\{x_1, x_2, \ldots, x_n\}$, these rules imply the normalization condition

$$\sum_{i=1}^{n} \mathbb{P}(x_i) = 1 \, ,$$

also referred to as *unitarity*, which simply states that *something must happen*.

The numerical value of a probability reflects how surprising an outcome would be. When $\mathbb{P}(A) \to 0$, the event is almost impossible; observing it would be highly surprising. When $\mathbb{P}(A) \to 1$, the event is almost certain; its occurrence carries little surprise. Between these extremes lies the full range of uncertainty, where probability quantifies degrees of expectation rather than absolute certainty or impossibility.

Why, for example, do we say that a fair coin has probability 1/2 of landing heads, or that a fair die has probability 1/6 of showing a given face? These numbers are not empirical facts but modeling assumptions based on symmetry. When all outcomes are assumed to be equally possible and indistinguishable before observation, probability assigns equal weight to each outcome. Probability theory then explores the logical consequences of these assumptions.

One common interpretation of probability is the *frequentist* view, developed most clearly by von Mises [38]. In this perspective, probability is identified with the long-run relative frequency of an event in repeated, identical experiments. Saying that a coin has probability 0.5 of landing heads means that, over many tosses, roughly half will result in heads.

An alternative interpretation is the *Bayesian* view, originating with Thomas Bayes [39] and developed further by Laplace and later authors such as de Finetti [40] and Jaynes [41]. Here,

probability quantifies uncertainty or degree of belief rather than long-run frequency. Probabilities are updated as new information becomes available using Bayes' theorem.

Both interpretations use the same mathematical rules and rely on Kolmogorov's axioms. The difference lies not in the calculations, but in how probability statements are interpreted. Again, Bayesian methods and their practical consequences will be introduced formally in 5.

2.2 Discrete events

By *discrete* we mean that the set of possible outcomes is finite or countably infinite. In such cases, probability distributions assign exact probabilities to individual outcomes and are therefore called probability *mass* functions. Discrete models are particularly useful when outcomes correspond to counts, successes and failures, or categorical observations. Bernoulli's work was motivated by a fundamental philosophical question: how can stable numerical regularities arise from individual events that appear completely unpredictable? His analysis of repeated trials provided one of the earliest mathematical explanations of how chance and regularity coexist.

Mathematically, a discrete probability distribution assigns a probability $\mathbb{P}(x_i)$ to each possible outcome x_i, such that

$$\mathbb{P}(x_i) \geq 0, \qquad \sum_{\forall i} \mathbb{P}(x_i) = 1. \tag{2.1}$$

where the $\forall$ symbol is just a mathematical character meaning "for all".

2.2.1 Bernoulli trials

The Bernoulli trial was formalized by Jacob Bernoulli in *"Ars Conjectandi"* (1713) [13]. His motivation was to understand how regularity emerges from randomness when an experiment with two outcomes is repeated many times.

A Bernoulli random variable X takes only two values, usually 1 (success) and 0 (failure). Its probability mass function is

$$\mathbb{P}(x;\ p) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0, \end{cases} \qquad 0 \leq p \leq 1. \tag{2.2}$$

Bernoulli trials are used whenever an experiment has exactly two possible outcomes. Typical examples include success or failure of a medical treatment, acceptance or rejection of a manufactured item, or whether a user clicks on a digital advertisement. In all these cases, the outcome is binary, even if the underlying process is complex.

Example. A single coin toss can be modeled as a Bernoulli trial, with $X = 1$ representing heads and $X = 0$ tails. For a fair coin, symmetry suggests $p = 1/2$.

2.2.2 Discrete uniform distribution

The discrete uniform distribution has its roots in classical symmetry arguments used in early probability theory. It formalizes the idea that, in the absence of distinguishing information, all outcomes should be treated equally. This idea reflects a principle already present in early probability theory: when no outcome can be distinguished from another based on available information, they should be treated symmetrically. Laplace formalized this reasoning as the principle of insufficient reason.

If a random variable X can take n distinct values $\{x_1, \ldots, x_n\}$, the discrete uniform distribution assigns

$$\mathbb{P}(x_i;\ n) = \frac{1}{n}, \qquad i = 1, \ldots, n. \tag{2.3}$$

(a) Bernoulli example 1.
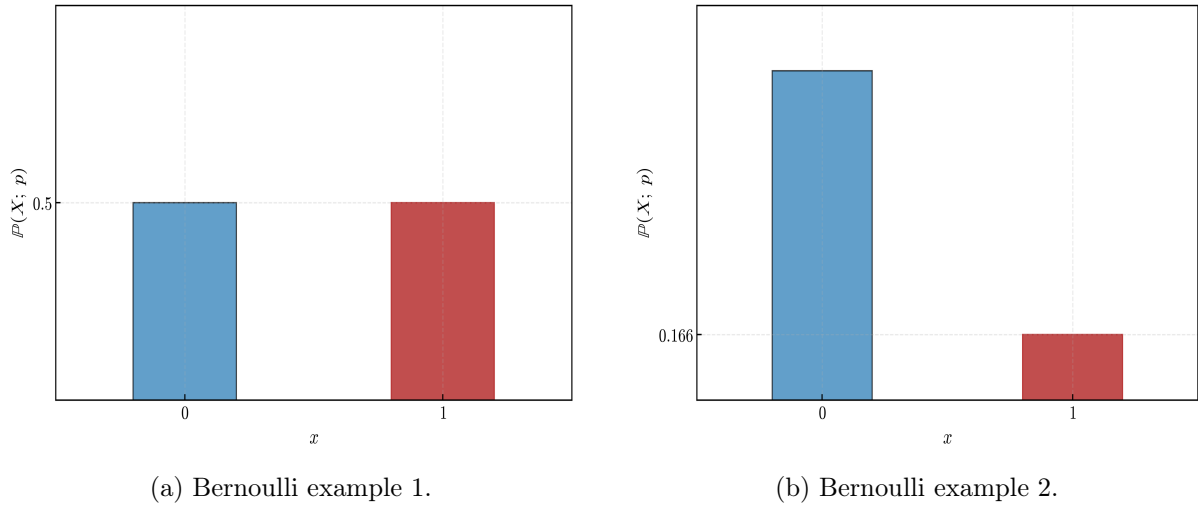
(b) Bernoulli example 2.

Figure 2.1: Probability distribution

Discrete uniform distributions appear whenever outcomes are assumed to be equally likely. Examples include lotteries, card draws from a well-shuffled deck, or randomized experimental assignments where each category is given equal probability.

Example. Rolling a fair six-sided die can be modeled as a discrete uniform distribution on $\{1, 2, 3, 4, 5, 6\}$, where each face has probability $1/6$.
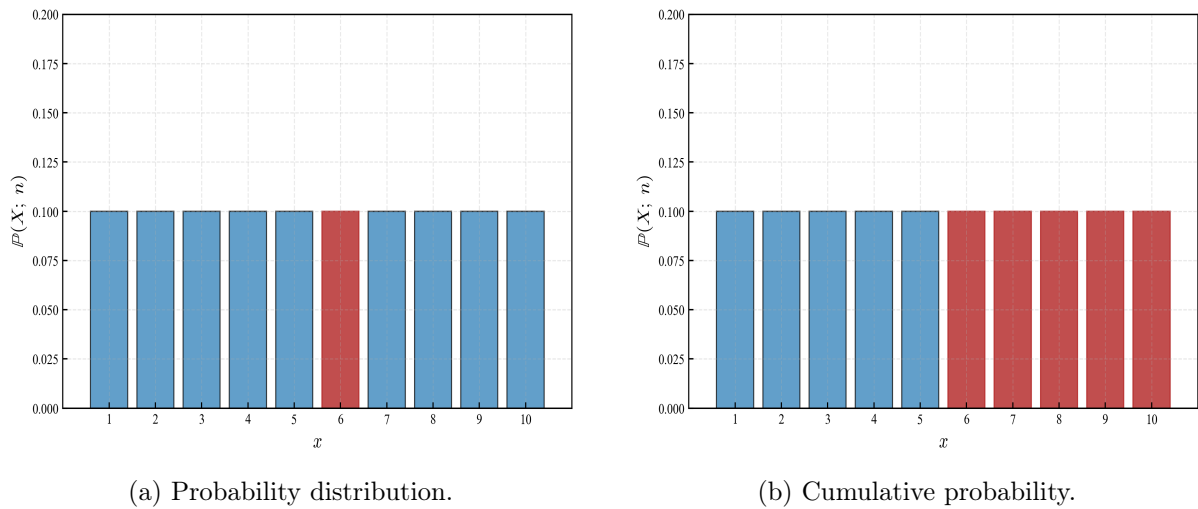

(a) Probability distribution.

(b) Cumulative probability.

Figure 2.2: Probability distribution

2.2.3 Binomial distribution

The binomial distribution was systematically studied by Abraham de Moivre in the early 18$^{\text{th}}$ century. His analysis of repeated Bernoulli trials led not only to the binomial formula but also to the first appearance of the normal approximation. De Moivre introduced the binomial distribution while studying games of chance, but its importance quickly extended far beyond gambling. By considering repeated trials under identical conditions, the binomial distribution became a central model for understanding variability in counting processes.

The binomial distribution models the number of successes in n independent Bernoulli trials

with success probability p. Its probability mass function is

$$\mathbb{P}(x;\ n,\ p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n. \tag{2.4}$$

Binomial models naturally arise when we count how many times a certain event occurs in a fixed number of attempts. Examples include the number of defective items in a batch, the number of patients responding to a treatment, or the number of voters favoring a candidate in a survey.

Example. The number of heads obtained when tossing a fair coin 10 times follows a binomial distribution with $n = 10$ and $p = 1/2$.
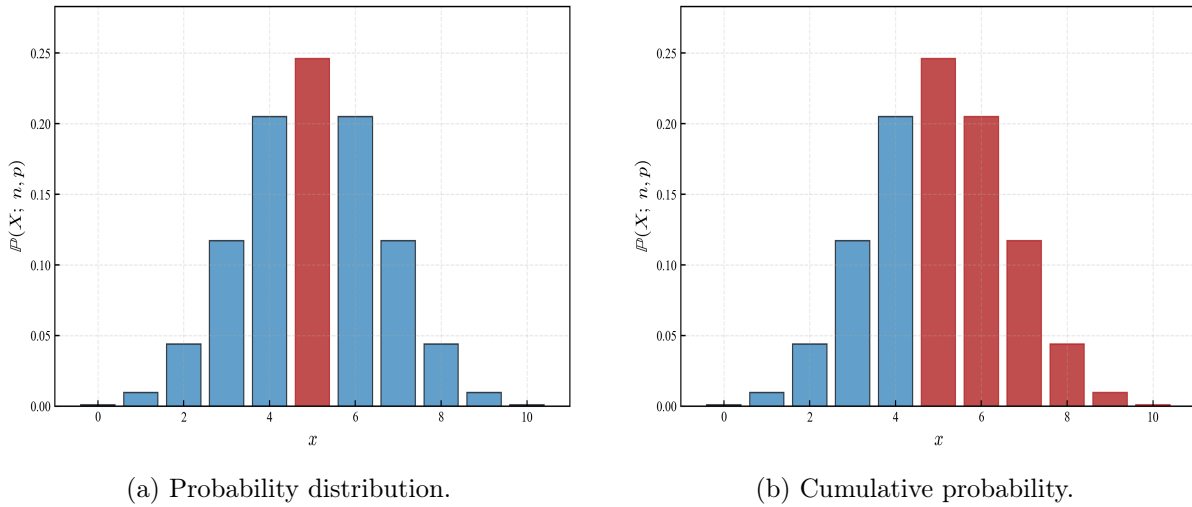

(a) Probability distribution.

(b) Cumulative probability.

Figure 2.3: Probability distribution

2.2.4 Poisson distribution

The Poisson distribution was introduced by Siméon Denis Poisson in 1837 while studying rare events in judicial statistics. It arises as a limiting case of the binomial distribution when events are rare but opportunities are numerous. Poisson originally introduced this distribution to study rare events, such as wrongful convictions in court cases. Its mathematical simplicity and clear interpretation soon made it a fundamental model for random counts occurring over time or space.

The Poisson distribution models the number of events occurring in a fixed interval of time or space. Its probability mass function is

$$\mathbb{P}(x;\ \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, 2, \ldots, \tag{2.5}$$

where $\lambda > 0$ is the average rate of occurrence.

Poisson distributions are commonly used to model events that occur independently and sporadically. Typical examples include the number of phone calls received by a call center, the number of typing errors on a page, or the number of decay events detected by a sensor during a fixed time interval.

Example. The number of emails received in one hour, when messages arrive independently at an average rate of $\lambda = 5$ per hour, can be modeled using a Poisson distribution.

Cumulative probabilities:

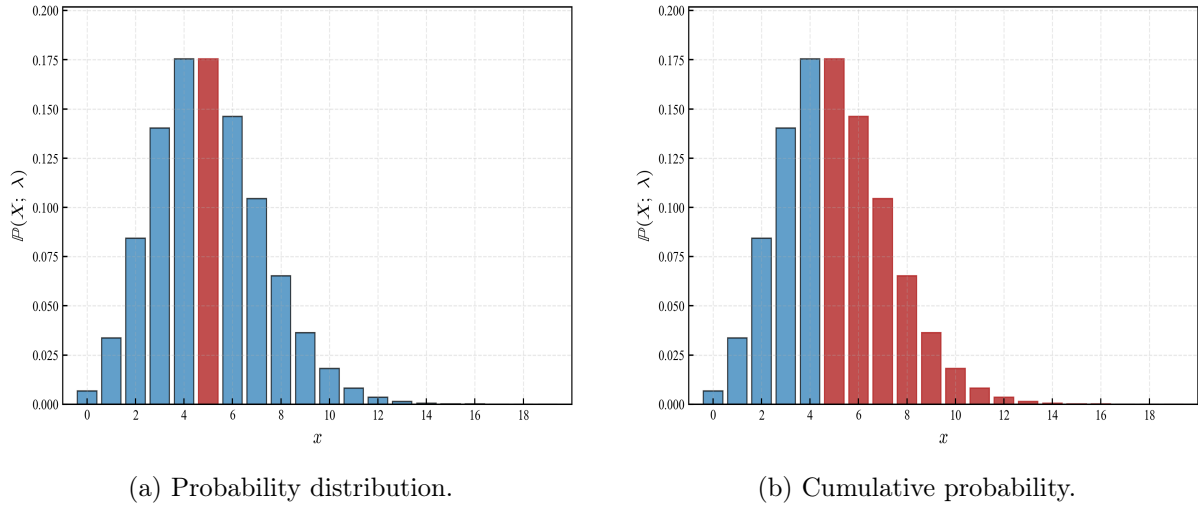(a) Probability distribution.   (b) Cumulative probability.

Figure 2.4: Probability distribution

So far, we have described probability distributions by assigning probabilities to individual outcomes. While this is natural for discrete models, it is often more informative to consider the probability that a random variable takes a value *up to* a given threshold. This leads to the notion of *cumulative probability*. As a note, it is common notation to use uppercase letters (such as X) to denote random variables, and lowercase letters (such as x) to denote their possible values.

For a discrete random variable X, the *cumulative distribution function* (CDF) is defined as

$$F(x) = \mathbb{P}(X \leq x), \tag{2.6}$$

that is, the probability that the outcome of the experiment does not exceed x.

If the possible values of X are $\{x_1, x_2, \dots\}$, the cumulative probability at a point x is obtained by summing the probabilities of all outcomes less than or equal to x:

$$F(x) = \sum_{x_i \leq x} \mathbb{P}(X = x_i). \tag{2.7}$$

The cumulative distribution function provides a global view of the distribution. While the probability mass function tells us how probability is assigned locally to each outcome, the CDF tells us how probability accumulates as we move along the real line.

Several important properties follow immediately from the definition:

- $F(x)$ is non-decreasing.

- $0 \leq F(x) \leq 1$ for all x.

- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.

Example. For a fair six-sided die, the probability that the outcome is at most 4 is

$$\mathbb{P}(X \leq 4) = \mathbb{P}(1) + \mathbb{P}(2) + \mathbb{P}(3) + \mathbb{P}(4) = \frac{4}{6}.$$

The corresponding cumulative distribution function increases in steps of size $1/6$ at each integer outcome.

Cumulative probabilities play a central role in probability theory because they naturally generalize to the continuous case, where individual outcomes no longer carry positive probability.

2.3   Continuous events

By *continuous* we mean that the set of possible outcomes is uncountably infinite, typically forming an interval of real numbers. In such cases, individual outcomes have zero probability, and uncertainty is described using probability *densities*. Probabilities are obtained by integrating the density over ranges of values.

When moving from discrete to continuous outcomes, the frequentist intuition that works well for counting events begins to break down. In a discrete setting, probabilities can be interpreted as long-run relative frequencies of individual outcomes. For example, the probability of rolling a 3 with a fair die can be understood as the fraction of times the outcome 3 appears when the die is rolled repeatedly. Each outcome has a positive probability, and frequencies converge to these values as the number of trials grows.

In a continuous setting, this interpretation can no longer be applied directly. If outcomes lie on a continuous interval, such as all real numbers between 0 and 1, the probability of observing any exact value is zero. No matter how many times the experiment is repeated, the relative frequency of obtaining exactly 0.37 will be zero. This does not mean that the outcome is impossible, but rather that probability must now be assigned to *ranges* of values rather than to individual points. The concept of a probability density is introduced precisely to resolve this issue: densities describe how probability is distributed locally, while actual probabilities are obtained by integrating the density over intervals. In this way, the frequentist idea of long-run relative frequency is preserved, but it applies to intervals of outcomes rather than to single values.

Mathematically, a continuous probability distribution is described by a density function $f(x)$ such that

$$f(x) \geq 0, \qquad \int_{-\infty}^{\infty} f(x)\, dx = 1. \tag{2.8}$$

The probability that a random variable lies in an interval $[a, b]$ is then given by the area under the density curve between a and b.

2.3.1   Gaussian distribution

The Gaussian distribution emerged from the work of Abraham de Moivre in the early 18th century and was later developed systematically by Pierre-Simon Laplace. Its physical interpretation was provided by Carl Friedrich Gauss in *Theoria Motus Corporum Coelestium* (1809) [15], in the context of measurement errors. Gauss introduced the distribution while studying astronomical observations, where repeated measurements of the same quantity produced small deviations around a central value. This interpretation linked probability theory directly to experimental science.

The Gaussian distribution models the accumulation of many small, independent effects. Its central role in probability theory is explained by the central limit theorem, which establishes it as a universal limiting distribution.

The probability density function of a Gaussian random variable X with mean μ and variance σ^2 is

$$f(x;\, \mu,\, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}. \tag{2.9}$$

Gaussian distributions are used to model many natural and social phenomena where values cluster around an average. Examples include measurement errors, biological traits such as height, and aggregated effects of many small influences acting together.

Example. Measurement errors in physical experiments are often modeled as Gaussian, with $\mu = 0$ representing no systematic bias and σ describing measurement precision.
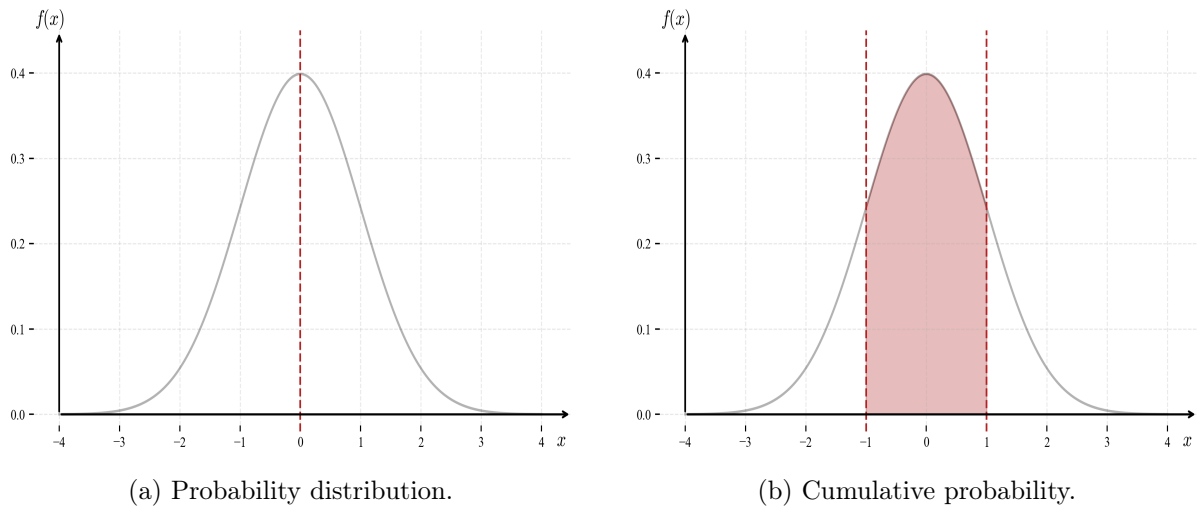
(a) Probability distribution.

(b) Cumulative probability.

Figure 2.5: Probability distribution

2.3.2   Exponential distribution

The exponential distribution arose in the 19th century in the study of waiting times and decay processes, closely connected to Poisson's work on random events and later developments in queueing theory [42, 20]. The exponential distribution emerged naturally from the study of random event timing, particularly in physics and telecommunications. Its mathematical form reflects the assumption that events occur independently and at a constant average rate.

It naturally models the time until the first occurrence of a random event and is characterized by the absence of memory: the future waiting time does not depend on how much time has already elapsed.

The probability density function of an exponential random variable X with rate $\lambda > 0$ is

$$f(x;\ \lambda) = \lambda e^{-\lambda x}, \qquad x \geq 0. \tag{2.10}$$

Exponential models are appropriate for waiting-time phenomena. Examples include the time until a machine fails, the time until the next customer arrives, or the time between successive radioactive decay events.

Example. The time until the next phone call arrives at a call center, assuming calls arrive independently at a constant average rate, is often modeled using an exponential distribution.

2.3.3   Continuous uniform distribution

The continuous uniform distribution extends classical symmetry arguments already present in Laplace's *Théorie Analytique des Probabilités* (1812) [16]. It represents complete ignorance about where within a bounded interval an outcome may fall. The continuous uniform distribution formalizes the idea of complete uncertainty over a bounded range. Unlike other distributions, it does not privilege any value within the interval, making it a neutral reference model.

If a random variable X is uniformly distributed on an interval $[a, b]$, its probability density function is

$$f(x;\ a,\ b) = \frac{1}{b-a}, \qquad a \leq x \leq b. \tag{2.11}$$

Continuous uniform distributions are often used in simulations and random sampling. They arise, for example, when generating random starting points, choosing random times within a fixed interval, or modeling unknown quantities constrained only by upper and lower bounds.
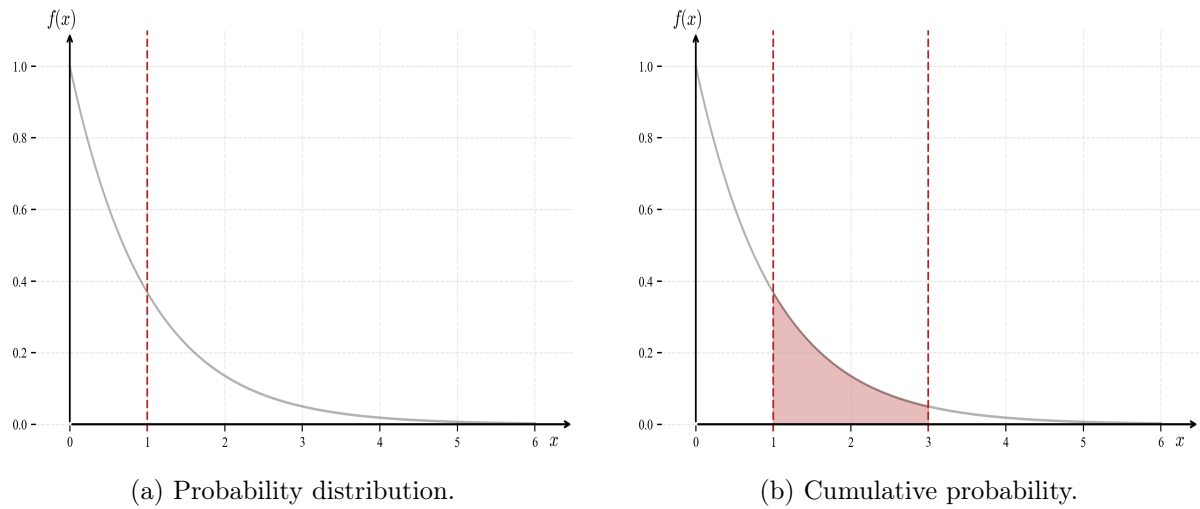
(a) Probability distribution.   (b) Cumulative probability.

Figure 2.6: Probability distribution

Example. If a random number generator produces values evenly between 0 and 1, the outcome can be modeled as a continuous uniform distribution on $[0, 1]$.
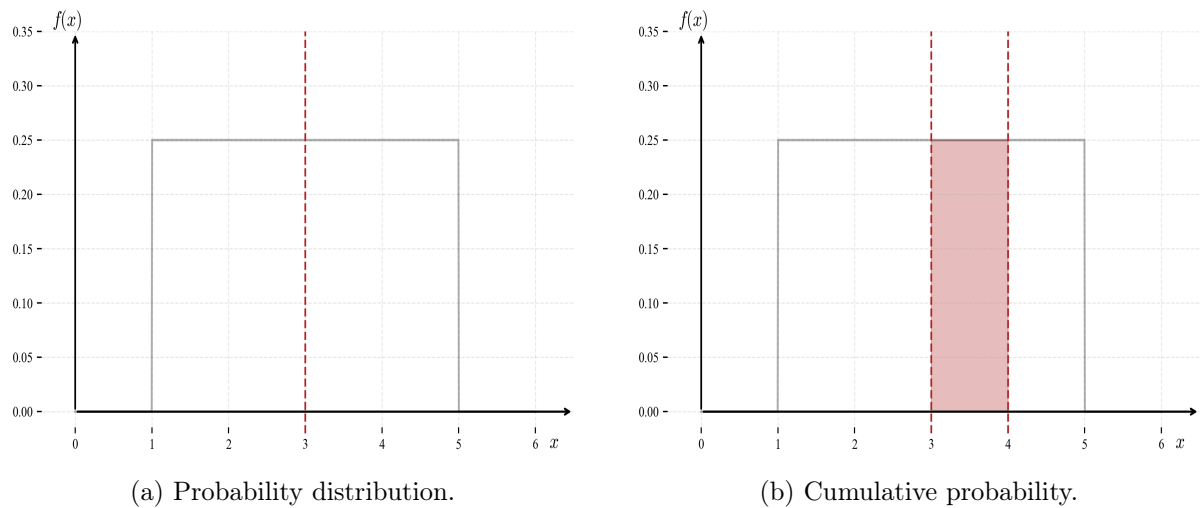

(a) Probability distribution.   (b) Cumulative probability.

Figure 2.7: Probability distribution

Cumulative probability in the continuous case:

When outcomes are continuous, individual values have zero probability and probability is described using densities rather than masses. Nevertheless, the idea of cumulative probability remains fundamental. As before, we use uppercase letters (such as X) to denote random variables, and lowercase letters (such as x) to denote their possible values.

For a continuous random variable X with probability density function $f(x)$, the cumulative distribution function is defined in exactly the same way:

$$F(x) = \mathbb{P}(X \leq x). \tag{2.12}$$

The difference lies in how this probability is computed. Since probability is distributed continuously, cumulative probability is obtained by integrating the density:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt. \tag{2.13}$$

In this setting, the density $f(x)$ does not represent probability itself, but rather the *rate at which probability accumulates*. The probability that X lies in an interval $[a, b]$ is given by

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(t) \, dt = F(b) - F(a). \tag{2.14}$$

Thus, while probability mass functions assign probabilities to individual points, probability density functions describe how probability is spread across the real line. In both discrete and continuous cases, the cumulative distribution function provides a unified description of uncertainty and serves as the primary object connecting probability to integration.

2.4 Expected values

Probability distributions describe uncertainty, but to summarize and compare them we often want a small number of representative quantities. The most important of these are the *moments* of a random variable. Moments capture different aspects of a distribution such as its location, spread, and shape.

Throughout this section, we will again denote random variables by uppercase letters, such as X. In the discrete case, probabilities are assigned to individual values via $\mathbb{P}(X = x_i)$, while in the continuous case probabilities are assigned to intervals and described through a probability density function.

The most fundamental moment is the *expected value*, also called the mean. Naming here may seem misleading, as this mean is not to be confused with the sample mean we discussed in previous chapter. One thing is the mean and variance *as expected values of random variables*, and a different one is as a number that summarizes a sample, which we referred in last chapter as *sample mean*. Informally, the expected value represents the long-run average outcome of a random experiment repeated many times. It answers the question: *where is the distribution centered?*

For a discrete random variable X taking values $\{x_i\}$ with probabilities $\mathbb{P}(X = x_i)$, the expected value is defined as

$$\mathbb{E}[X] = \sum_i x_i \, \mathbb{P}(X = x_i). \tag{2.15}$$

For a continuous random variable with probability density function $f(x)$, the expected value is defined analogously by replacing the sum with an integral:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \, f(x) \, dx. \tag{2.16}$$

The expected value is a *first-moment* quantity: it captures the location of a distribution but provides no information about its variability. A key property of expectation is linearity. For any random variables X_i,

$$\mathbb{E}\left[\sum_i X_i\right] = \sum_i \mathbb{E}[X_i], \tag{2.17}$$

regardless of whether the variables are independent.

The second moment of central importance is the *variance*, which measures how spread out the distribution is around its mean. Variance is defined as the expected squared deviation from the mean:

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]. \tag{2.18}$$

An equivalent and often more convenient expression for the variance is

$$\text{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2. \tag{2.19}$$

For a discrete random variable, this corresponds to

$$\text{Var}(X) = \sum_i (x_i - \mu)^2 \, \mathbb{P}(X = x_i), \qquad \mu = \mathbb{E}[X], \tag{2.20}$$

while for a continuous random variable it is given by

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \, f(x) \, dx. \tag{2.21}$$

More generally, the k-th *moment* of a random variable describes higher-order features of its distribution:

- The *first moment* (mean) describes location.

- The *second moment* (variance) describes spread.

- The *third moment* is related to *skewness*, measuring asymmetry.

- The *fourth moment* is related to *kurtosis*, measuring tail heaviness.

In practice, mean and variance already provide a powerful summary of most distributions. Classical statistical inference—such as confidence intervals, hypothesis tests, and error propagation—relies heavily on estimators of these two quantities and on their sampling distributions.

2.5   Mean and variance of common distributions

We conclude this chapter by collecting the expected value and variance of the probability distributions introduced earlier. These results provide concrete examples of the abstract definitions given in the previous section and will be used repeatedly in later chapters.

Bernoulli distribution. Let $X \sim \text{Bern}(p)$, with $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$. Then

$$\mathbb{E}[X] = p, \tag{2.22}$$

$$\text{Var}(X) = p(1 - p). \tag{2.23}$$

The mean equals the success probability, while the variance is largest when $p = 1/2$.

Binomial distribution. Let $X \sim \text{Bin}(n, p)$ represent the number of successes in n independent Bernoulli trials. Then

$$\mathbb{E}[X] = np, \tag{2.24}$$

$$\text{Var}(X) = np(1 - p). \tag{2.25}$$

Both the mean and variance scale linearly with the number of trials.

Poisson distribution. Let $X \sim \text{Pois}(\lambda)$, where $\lambda > 0$ is the average rate of occurrence. Then

$$\mathbb{E}[X] = \lambda, \tag{2.26}$$

$$\text{Var}(X) = \lambda. \tag{2.27}$$

A defining feature of the Poisson distribution is that its mean and variance coincide.

Discrete uniform distribution. Let X be uniformly distributed on the set $\{1, 2, \ldots, n\}$. Then

$$\mathbb{E}[X] = \frac{n+1}{2}, \tag{2.28}$$

$$\mathrm{Var}(X) = \frac{n^2 - 1}{12}. \tag{2.29}$$

The mean lies at the center of the interval, while the variance depends only on its width.

Gaussian distribution. Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\mathbb{E}[X] = \mu, \tag{2.30}$$

$$\mathrm{Var}(X) = \sigma^2. \tag{2.31}$$

The parameters μ and σ^2 directly control the location and spread of the distribution.

Exponential distribution. Let $X \sim \mathrm{Exp}(\lambda)$, with $\lambda > 0$. Then

$$\mathbb{E}[X] = \frac{1}{\lambda}, \tag{2.32}$$

$$\mathrm{Var}(X) = \frac{1}{\lambda^2}. \tag{2.33}$$

Larger values of λ correspond to shorter expected waiting times and reduced variability.

Continuous uniform distribution. Let $X \sim \mathrm{Unif}(a, b)$. Then

$$\mathbb{E}[X] = \frac{a+b}{2}, \tag{2.34}$$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}. \tag{2.35}$$

As in the discrete case, the mean lies at the midpoint of the interval and the variance depends only on its length.

These examples illustrate how mean and variance summarize probability distributions in concrete terms. In later chapters, we will study how these quantities are estimated from data and how their sampling variability affects statistical inference.

Chapter 3

Estimation, variability and confidence

The development of mathematics is a continuous process of abstraction.

— Emmy Noether

In the previous chapters, we described data using summary statistics and introduced probability models to represent uncertainty. In this chapter, we connect these two perspectives. Estimation is the process through which data are used to learn about unknown features of a population, while probability provides the language to quantify how reliable such learning is.

As we discussed already, a crucial distinction must be made between *prediction* and *inference*. Prediction focuses on forecasting future observations, whereas inference aims to draw conclusions about underlying parameters or mechanisms that generate the data. Estimation lies at the heart of inference: it transforms random samples into numerical statements about unknown quantities.

Because data are inherently variable, different samples lead to different estimates. Understanding how estimates behave across repeated samples is therefore essential. The main goal of this chapter is to explain how probability theory allows us to quantify this variability and to construct principled measures of uncertainty such as confidence intervals.

3.1   The Law of Large Numbers

Historically, the Law of Large Numbers was first proved by Jacob Bernoulli in his posthumously published *Ars Conjectandi* (1713). Bernoulli's motivation was not purely mathematical: he sought to justify how stable numerical patterns could emerge from seemingly random individual events. His theorem provided the first rigorous foundation for interpreting probability in terms of long-run relative frequencies and marked a decisive step in connecting abstract probability with empirical observation [13, 14].

The Law of Large Numbers formalizes the intuitive idea that averages stabilize when more data are collected. While individual observations may be highly variable, their average becomes increasingly predictable as the sample size grows.

Historically, this result was first articulated by Jacob Bernoulli in the early eighteenth century. It provided the first rigorous justification for interpreting probability as long-run relative frequency and established a bridge between theoretical probability and empirical observation.

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with expected value μ. The Law of Large Numbers states that the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{3.1}$$

converges to μ as the sample size n increases.

Example. When repeatedly tossing a fair coin, individual outcomes are unpredictable, but the proportion of heads approaches $1/2$ as the number of tosses grows.
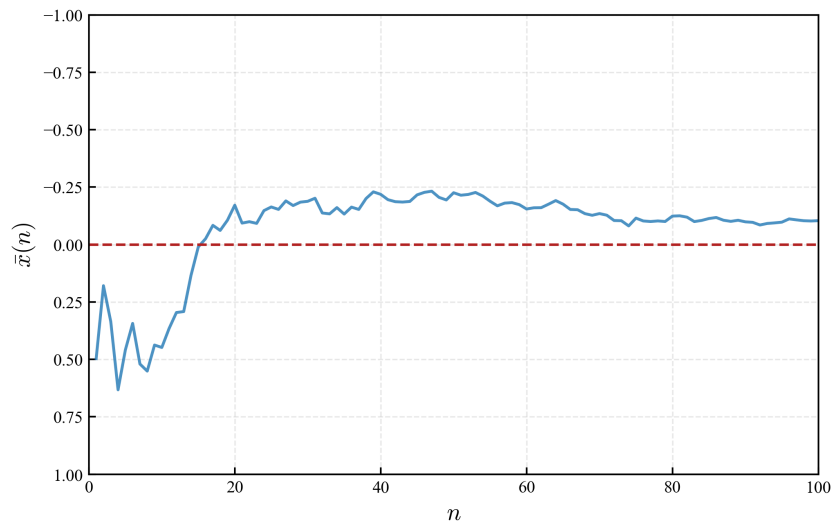

Figure 3.1: The Law of Large Numbers.

3.2 The Central Limit Theorem

The origins of the Central Limit Theorem can be traced to the work of Abraham de Moivre in the early eighteenth century, who discovered that binomial distributions with large numbers of trials could be approximated by a normal curve. This insight was later generalized by Pierre-Simon Laplace, who showed that the normal distribution arises broadly from the aggregation of many independent random effects. The modern formulation of the theorem, with precise conditions and convergence statements, was completed only in the late nineteenth and early twentieth centuries [16, 24].

While the Law of Large Numbers explains where averages converge, it does not describe how they fluctuate around their limiting value. The Central Limit Theorem answers this question by describing the distribution of fluctuations.

One of the most remarkable results in probability theory, the Central Limit Theorem explains why the Gaussian distribution appears so frequently in statistical practice. It shows that many different random processes give rise to approximately normal behavior when aggregated.

Let $X_1, \ldots, X_n$ be independent and identically distributed with mean μ and variance σ^2. The Central Limit Theorem states that the standardized sample mean

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \tag{3.2}$$

approaches a standard normal distribution as n becomes large.

Example. Even if individual measurements are skewed, the distribution of their average over many observations is often approximately Gaussian.

3.3 Bias, variance and Mean Squared Error

The systematic study of estimators and their properties developed alongside the rise of mathematical statistics in the late nineteenth and early twentieth centuries. Early statisticians recognized that good estimation procedures must balance systematic error and random variability. This insight led to the decomposition of mean squared error into variance and squared bias, a framework that remains central in both classical statistics and modern machine learning [24, 3].

An estimator is a rule that assigns a numerical value to an unknown parameter based on observed data. Because estimators depend on random samples, they are themselves random variables.

The *bias* of an estimator measures systematic error: it quantifies whether the estimator tends to overestimate or underestimate the true parameter. The *variance* measures how much the estimator fluctuates from sample to sample.

If $\hat{\theta}$ is an estimator of a parameter θ, its bias is defined as

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta, \tag{3.3}$$

and its variance is

$$\text{Var}(\hat{\theta}). \tag{3.4}$$

A common way to combine these two sources of error is the mean squared error (MSE):

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2. \tag{3.5}$$

Example. The sample mean is an unbiased estimator of the population mean. Increasing the sample size reduces its variance, making the estimate more precise.

3.4 Confidence intervals and critical regions

Confidence intervals emerged as a practical response to the limitations of point estimation. Rather than asking for a single best value, statisticians sought procedures that quantify uncertainty in a repeatable and operational way. The concept was formalized in the early twentieth century as part of the frequentist framework, emphasizing long-run coverage properties under repeated sampling. This interpretation remains foundational in applied statistics and experimental science [3, 4].

Point estimates summarize data with a single number, but they do not convey uncertainty. Confidence intervals address this limitation by providing a range of plausible values for an unknown parameter. A confidence interval is constructed so that, in repeated sampling, it contains the true parameter a fixed proportion of the time. This proportion is called the confidence level and is typically expressed as a percentage.

For a population with mean μ and known variance σ^2, an approximate $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \tag{3.6}$$

where $z_{\alpha/2}$ is a quantile of the standard normal distribution.

Confidence intervals are closely related to hypothesis testing. The set of parameter values not rejected by a statistical test forms a confidence region. This duality provides a unified framework for estimation and decision-making.
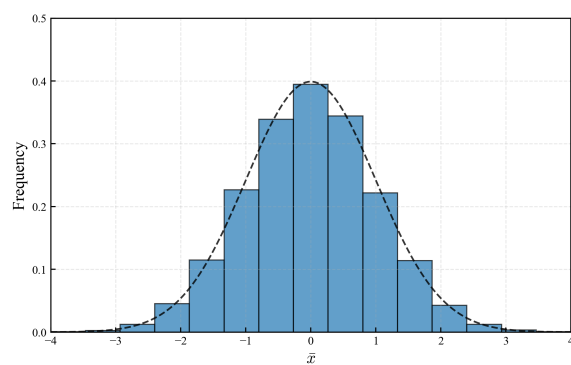
Example. A 95% confidence interval for the mean expresses the range of values that are consistent with the observed data under repeated sampling.
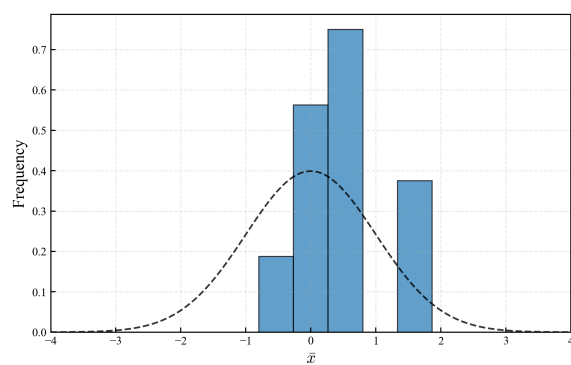
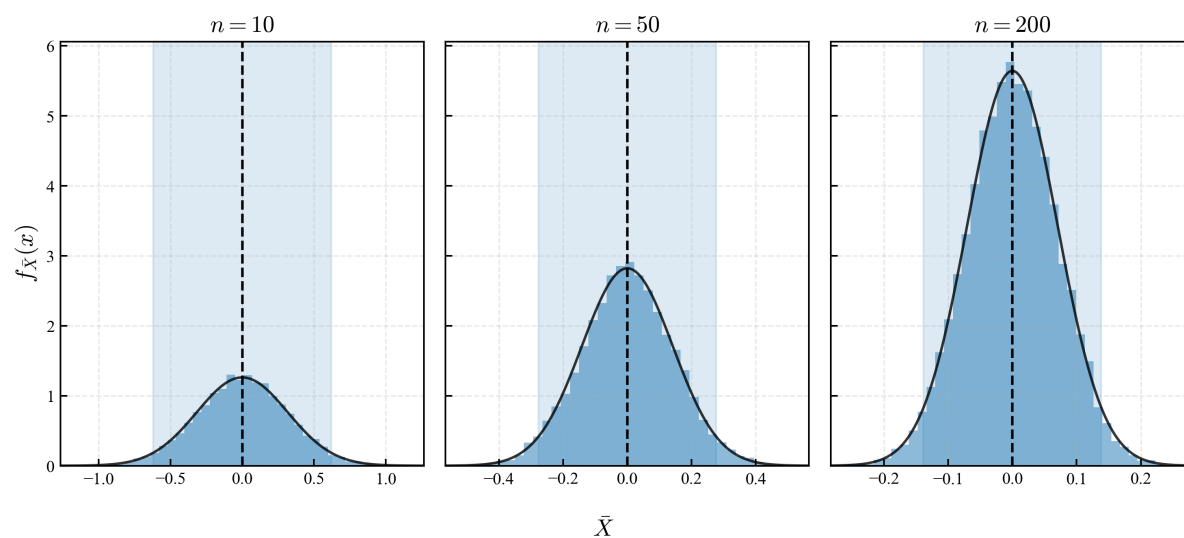(a) The Central Limit Theorem 1
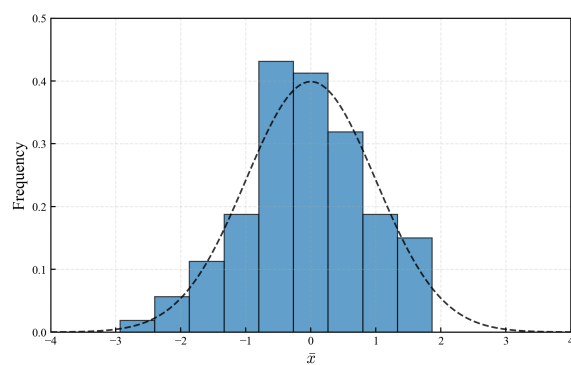
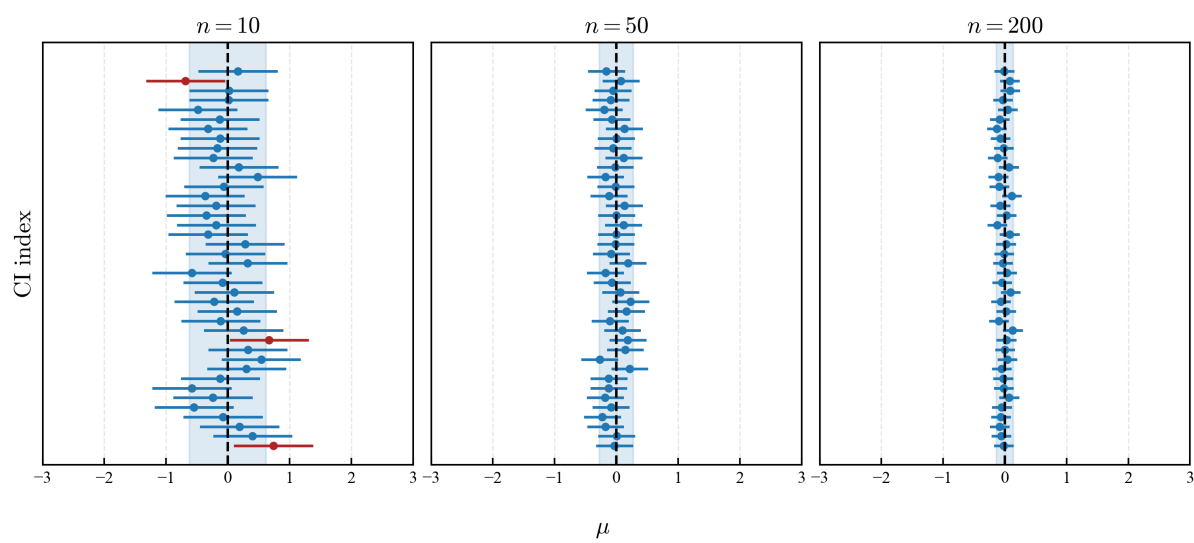(b) The Central Limit Theorem 2

(c) The Central Limit Theorem 3

Figure 3.2: The Central Limit Theorem

(a) Sample mean distribution

(b) Confidence intervals

Figure 3.3: The Central Limit Theorem

Chapter 4

Introduction to hypothesis testing

Chapter 5

Modelling, dependency, and correlation

Chapter 6

Introduction to Bayesian probability

Chapter 7

Stochasticity and Markov Processes

Appendix A

Linear algebra: core topics on vectors and matrices

Linear algebra is one of the central languages of modern mathematics and science. It provides the formal framework for describing linear relations, symmetry, and structure, and it underlies vast areas of analysis, probability, statistics, physics, computer science, and data science. Historically, linear algebra did not arise as a single theory, but rather as a collection of methods developed to solve systems of linear equations, study geometry, and understand transformations of space. Only in the late nineteenth and early twentieth centuries was it unified into the abstract theory now taught under the name *linear algebra*.

From a philosophical perspective, linear algebra marks a transition from concrete computation to structural thinking. Instead of focusing on individual equations or numerical solutions, the subject emphasizes vector spaces, mappings between them, and invariant properties under change of coordinates. This abstraction allows the same mathematical ideas to apply equally to geometry, differential equations, probability models, and statistical inference.

A solid background in linear algebra typically includes the following core topics, which are either required or strongly recommended for further study in probability, statistics, and applied mathematics:

Vectors and vector spaces. Understanding vectors as elements of a vector space over a field, together with the operations of addition and scalar multiplication. This includes linear combinations, span, linear independence, bases, and dimension. These concepts formalize the idea of degrees of freedom and coordinate representations.

Matrices and systems of linear equations. Matrices arise naturally as representations of linear maps. Key topics include matrix operations, matrix inversion, rank, row reduction, and the solution of linear systems. Historically, these ideas trace back to the work of Carl Friedrich Gauss and the development of systematic elimination methods.

Linear transformations. The interpretation of matrices as functions between vector spaces is essential. Topics include kernel and image, injectivity and surjectivity, change of basis, and composition of linear maps. This viewpoint emphasizes structure over computation and clarifies the geometric meaning of matrices.

Inner products and geometry. Inner product spaces introduce notions of length, angle, and orthogonality. Important concepts include norms, orthogonal projections, orthonormal bases, and the Gram–Schmidt process. These ideas are foundational for least squares methods, regression, and statistical estimation.

Eigenvalues and eigenvectors. Eigenvalues describe intrinsic directions and scaling properties of linear transformations. Diagonalization, spectral decomposition, and symmetric matrices

play a central role in applications ranging from differential equations to principal component analysis.

Determinants and volume. Although less central in modern abstract treatments, determinants provide geometric insight into volume, orientation, and invertibility. They also play a role in change-of-variables formulas and multivariate analysis.

Historically, the development of linear algebra is associated with several key figures and works. Gauss laid the foundations for systematic solution of linear systems. Hermann Grassmann introduced abstract vector spaces in his *Ausdehnungslehre* (1844), a work far ahead of its time. Arthur Cayley and James Joseph Sylvester developed matrix theory in the nineteenth century, while David Hilbert and Emmy Noether contributed decisively to the structural and axiomatic understanding of linear spaces and linear operators.

In modern mathematics, linear algebra serves both as a computational toolkit and as a conceptual foundation. Mastery of its basic structures is essential not only for solving concrete problems, but also for understanding more advanced theories where linear spaces provide local or approximate descriptions of complex phenomena. As such, linear algebra is best learned not merely as a collection of techniques, but as a coherent framework for reasoning about structure, symmetry, and linearity.

Appendix B

Calculus I: Core topics on functions and derivatives

Calculus begins with the study of how quantities depend on one another and how they change. At its foundation lies the concept of a *function*, which formalizes the idea that one quantity is determined by another. Functions provide the language through which variation, motion, and growth are described in mathematics, physics, and the natural sciences.

Historically, the notion of a function evolved gradually. Early uses appear implicitly in the work of René Descartes, who introduced coordinate geometry and expressed curves through algebraic equations. The explicit concept of a function as a mapping between quantities was later clarified in the eighteenth century by Leonhard Euler, whose writings established much of the notation and terminology still in use today.

A central idea in calculus is that of a *limit*. Limits capture the behavior of a function as its input approaches a given value, even if the function is not defined or not well behaved at that point. Informally, limits allow us to reason about processes that involve approaching, rather than reaching, a value. This concept is essential for making precise sense of continuity, instantaneous change, and accumulation.

The derivative arises from the study of limits and provides a precise definition of instantaneous rate of change. Geometrically, the derivative of a function at a point corresponds to the slope of the tangent line at that point. Physically, it describes quantities such as velocity or growth rate. The basic definition of the derivative is given by a limit of difference quotients, linking algebraic computation with geometric intuition.

The development of differential calculus is closely associated with Isaac Newton and Gottfried Wilhelm Leibniz, who independently formulated its fundamental principles in the late seventeenth century. Newton emphasized motion and physical interpretation, while Leibniz introduced a symbolic notation that proved especially flexible and influential. Their work laid the groundwork for centuries of mathematical and scientific progress.

From a philosophical standpoint, calculus represents an effort to make sense of continuous change using finite reasoning. The introduction of limits resolved long-standing paradoxes about infinity and infinitesimals by replacing informal arguments with precise definitions. Modern calculus, as taught today, builds on this foundation by emphasizing clarity, rigor, and conceptual understanding rather than purely mechanical computation.

Appendix C

Calculus II: Core topics on integral

Integral calculus is concerned with accumulation, total change, and the measurement of quantities such as area, volume, and total mass. Whereas differential calculus studies how quantities change at a point, integral calculus focuses on how these changes add up over an interval. Together, the two form a unified framework for analyzing continuous phenomena.

The basic problem of integration can be stated simply: given a varying quantity, how can one compute its total effect? Early examples include determining the area under a curve or the distance traveled by an object with varying speed. These questions were studied in ancient mathematics, notably by Archimedes, who used geometric arguments to compute areas and volumes with remarkable precision.

In modern terms, the integral of a function over an interval is defined as the limit of finite sums, where the interval is subdivided and the contributions of each subinterval are added together. This idea formalizes the intuitive notion of accumulation and provides a bridge between discrete approximation and continuous exactness.

A fundamental result of calculus is the *Fundamental Theorem of Calculus*, which establishes a deep connection between differentiation and integration. It shows that integration can be performed by finding an antiderivative, linking the problem of accumulation directly to the study of rates of change. This theorem unifies the two branches of calculus into a single coherent theory.

As with differential calculus, the systematic development of integral calculus is attributed to Newton and Leibniz. Leibniz's notation for integrals, inspired by the idea of summation, remains standard today. Over time, the theory of integration was refined and extended by mathematicians such as Augustin-Louis Cauchy and Bernhard Riemann, who provided precise definitions suitable for rigorous analysis.

Philosophically, integral calculus addresses the challenge of understanding the whole from infinitely many parts. By replacing informal geometric reasoning with limit processes, it provides a reliable method for reasoning about continuous quantities. At an introductory level, integral calculus offers both practical tools and conceptual insight into how local behavior combines to produce global effects.

Bibliography

[1] John P. A. Ioannidis. "*Why Most Published Research Findings Are False*". *PLoS Medicine*, 2(8):e124, 2005.

[2] David Spiegelhalter. *"The Art of Statistics: How to Learn from Data"*. Basic Books, 2019.

[3] Morris H. DeGroot and Mark J. Schervish. *"Probability and Statistics"*. Pearson, 4 edition, 2012.

[4] P. S. Bandyopadhyay and M. R. Forster, editors. *"Philosophy of Statistics"*, volume 7 of *Handbook of the Philosophy of Science*. Elsevier, 2011.

[5] M. Diez, D. Barr, and Mine Çetinkaya-Rundel. *"OpenIntro Statistics"*. OpenIntro, 2025.

[6] Hossein Pishro-Nik. *"Introduction to Probability, Statistics and Random Processes"*. Kappa Research LLC, 2014.

[7] Irving L. Finkel. "*The Ancient Origins of Dice*". *Antiquity*, 81(314):176–187, 2007.

[8] F. N. David. *"Games, Gods and Gambling"*. Griffin, 1962.

[9] Marcus Tullius Cicero. *"De Divinatione (On Divination"*. Ancient Sources Edition, 45 BCE.

[10] Gerolamo Cardano. *"Liber de Ludo Aleae (Book on Games of Chance)"*. Apud Joannem Baptistam Ferrarium, Paris, 1663.

[11] Keith Devlin. *"The Unfinished Game: Pascal, Fermat, and the seventeenth-century letter that made the world modern"*. Basic Books, 2008.

[12] Christiaan Huygens. *"De Ratiociniis in Ludo Aleae (On Reasoning in Games of Chance)"*. Elzevier, Leiden, 1657.

[13] Jacob Bernoulli. *"Ars Conjectandi (The Art of Conjecturing)"*. Thurneysen Brothers, Basel, 1713.

[14] Anders Hald. *"A History of Probability and Statistics and Their Applications before 1750"*. Wiley, 1990.

[15] Carl Friedrich Gauss. *"Theoria motus corporum coelestium in sectionibus conicis solem ambientium"*. Perthes et Besser, Hamburg, 1809.

[16] Pierre-Simon Laplace. *"Théorie Analytique des Probabilités"*. Courcier, Paris, 1812.

[17] Adolphe Quetelet. *"Sur l'homme et le développement de ses facultés, ou essai de physique sociale"*. Bachelier, Paris, 1835.

[18] Florence Nightingale. *"Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army"*. Harrison and Sons, London, 1858.

[19] Andrey Andreyevich Markov. *"Rasprostranenie zakona bol'shikh chisel na velichiny, zavisyashchie drug ot druga (Extension of the Law of Large Numbers to Dependent Variables)"*. Imperatorskaya Akademiya Nauk, St. Petersburg, 1906.

[20] Joseph L. Doob. *"Stochastic Processes"*. John Wiley & Sons, New York, 1953.

[21] Andrey Kolmogorov. *"Grundbegriffe der Wahrscheinlichkeitsrechnung (Foundations on the Theory of Probability)"*. Springer, Berlin, 1933.

[22] Wesley C. Salmon. *"Scientific Explanation and the Causal Structure of the World"*. Princeton University Press, Princeton, 1984.

[23] Clark Glymour. *"Theory and Evidence"*. Princeton University Press, Princeton, 1980.

[24] Stephen M. Stigler. *"The History of Statistics: The Measurement of Uncertainty before 1900"*. Harvard University Press, Cambridge, MA, 1986.

[25] Ian Hacking. *"The Emergence of Probability"*. Cambridge University Press, Cambridge, 1975.

[26] Frank P. Ramsey. *"The Foundations of Mathematics and Other Logical Essays"*. Routledge & Kegan Paul, London, 1931.

[27] Karl Pearson. *"The Grammar of Science"*. Adam and Charles Black, London, 1892. Foundational work on statistical reasoning and frequency distributions.

[28] Karl Pearson. *"Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material"*. *Philosophical Transactions of the Royal Society of London A*, 186:343–414, 1895.

[29] John W. Tukey. *"Exploratory Data Analysis"*. Addison-Wesley, 1977.

[30] J. L. Hintze and R. D. Nelson. *"Violin Plots: A Box Plot-Density Trace Synergism"*. *The American Statistician*, 52(2):181–184, 1997.

[31] Charles Spearman. *"The Proof and Measurement of Association Between Two Things"*. *The American Journal of Psychology*, 15(1):72–101, 1904.

[32] F. J. Anscombe. *"Graphs in Statistical Analysis"*. *The American Statistician*, 27(1):17–21, 1973.

[33] René Descartes. *"La Géométrie"*. Jan Maire, Leiden, 1637.

[34] Charles Joseph Minard. *"Tableau Graphique des Pertes Successives en Hommes de l'Armée Française dans la Campagne de Russie 1812–1813"*. Imprimerie Nationale, Paris, 1862.

[35] William Playfair. *"The Commercial and Political Atlas"*. J. Debrett, London, 1786.

[36] Henri Lebesgue. *"Intégrale, longueur, aire"*. 1902.

[37] Patrick Billingsley. *"Probability and Measure"*. Wiley, 1995.

[38] Richard von Mises. *"Probability, Statistics and Truth"*. 1928.

[39] Thomas Bayes. "An Essay towards Solving a Problem in the Doctrine of Chances". *Philosophical Transactions of the Royal Society of London*, 1763.

[40] Bruno de Finetti. *Theory of Probability*. Wiley, 1974.

[41] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[42] Siméon-Denis Poisson. *"Recherches sur la probabilité des jugements"*. 1837.