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

Jesús Urtasun Elizari

January 27, 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 five 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 to games of change 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 [...]

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. Thinkers such as Cicero 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"* [12], and Jacob Bernoulli, whose 1713 *"Ars Conjectandi"* 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].

It is from the 19th century onwards, that probability theory began to intertwine with statistics and inference, building the modern mathematical frameworks that we use nowadays to analyze and model physical phenomena. Florence Nightingale, best known for her pioneering role in modern nursing, made significant contributions to statistical methodology and graphical representation of data. Her advocacy for statistical reasoning in public health policy helped popularize quantitative approaches to uncertainty and variation. Around the same period, Joseph Fourier's work on heat conduction introduced Fourier series and integral transforms, tools that would later become indispensable for studying random processes, including the analysis of signals, noise, and diffusion phenomena. Although Nightingale and Fourier approached problems of uncertainty from very different perspectives—one through empirical data on human wellbeing, the other through mathematical physics—their contributions expanded the reach of probabilistic thinking and prepared the ground for future developments in stochastic analysis. [...]

A further conceptual leap, worth mentiong, occurred in the early 20th century with the work of Andrey Markov. Motivated partly by a desire to extend the law of large numbers beyond the assumption of independent trials, Markov developed what are now known as Markov chains, thereby inaugurating the study of dependence structures in stochastic processes. His investigations demonstrated that long-run statistical regularities could emerge even when successive events were not independent, a discovery that profoundly influenced both theoretical probability and its applications in fields as diverse as statistical mechanics, linguistics, quantum mechanics, and modern machine learning.

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"*) [15], 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 Chapter 5. 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 probabilty and statistics, we should properly define both topics and their main fields of study. Even deeply related, and both rooted in *combinatorics*—the study of uncertainty and things that 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, also referred to as *stochastic*, aiming to compute expected values for such unknown outcomes. On the other hand, statistics would be a *descriptive* way of dealing with uncertainty, by sampling finite sets of observations from a given population, and building informative quantities, called statistical *estimators* to explore central tendency and variation. Such distinction has been extensively debated and discussed by mathematicians, experimental scientists, and philosoplers of science.

As a rule of thumb, probability provides a formal language for modelling uncertainty, whereas statistics concerns the epistemic problem of learning from data. Through this chapter we will introduce basic ideas on statistical inference such as population, sampling, and estimators of central tendency and variation, together with some 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. Hence, 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 who encounter uncertainty in their work. In practice, there is a restriction in that statistics is ordinarily associated with data; and it is the link between the uncertainty, or variability, in the data and that in the topic itself that has occupied statisticians. Statistics does not have a monopoly of studies of uncertainty. Probability discusses how randomness in one part of a system affects other parts.

As a note, let us emphasize how these two approaches can and do coexist in science. We have many times heard that science works by making hypothesis and then predictions, that are compared and bechmarked with an experiment. This is a simplification, and it is not always true. Some sciences, like Newtonian mechanics, most of physics, chemistry, and certainly parts of biology, do rely on building accurate models and predictions, that are later compared with an experimental result. A clear example would be to use Newtonian mechanics as our theory, or model, to compute a prediction on where and when would a stone fall if I throw it. Then the experiment would be simply to measure, when and where. On the other hand, the archetypical example of an inference problem, which does not aim to build a prediction, but to give—or *reconstruct* or *infer*—an explanation given a set of observations, would be Darwinian evolution. This distinction is worth mentioning, since the usual definitions of sciences tend to rely heavily on the predictive power, which can be inaccurate and misleading [...]. Different sciences may strongly differ on methods, instrumentation, or conceptual tools, but they are all equally legitimate, regarltends to be defined

1.1 Sampling and data types

A large part of history of science could be summarized as a continous effort to translate observations of reality into precise, mathematical terms. To such endeavour, of describing the vast phenomena we find in the natural world with numerical language, it is necessary to develop tools that relate the one or more relevant quantities—sometimes called *variables*—and how they relate or change depending on one another. The purpose of modelling might be, for instance, to determine the distance from the earth to the sun, estimate the number of stars in the observable universe, or relate the number of lung cancer patients to pollution levels around smoking areas.

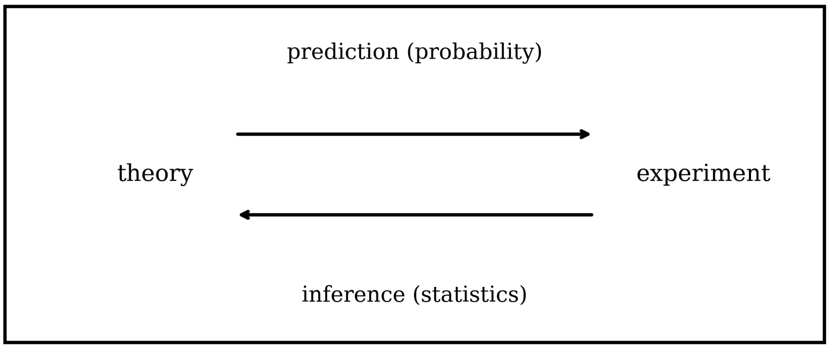


Figure 1.1: Representation of the predictive (from theory, or model, to experimental verification) and inferential (from data, measurement, observation to underlying truth) approaches to natural phenomena. As an example of the predictive branch of mathematics dealing with uncertainty we would find the theory of probability, while the descriptive way of addressing the same problem is normally regarded as statistical inference.

In the same way mathematics as a whole has been summarized as three simple tasks—*count*, *measure*, and *sort*—we could group statistical problems in three main groups. The problem of *sampling*—selecting a finite group of observations from a larger, unknown population— the *estimation*—build some mathematical quantity that represents how the measurements of my sample are distributed, and finally *visualization*—how my observations look like, and how that changes if I represent them in one way or another. Again, all of these problems are related to the phenomenon of *uncertainty*, or *variation* among measurements.

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 overflows with phenomena, offering 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

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

where the elements the sample x_i consist of just a selected group of observations from the population, not necessarily consecutive or in the same order. 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, I will never have access to the *complete population*, but only the amout 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 been associated with games of chance and gambling from quite old times, but it was not adressed as a statistics problem until much later. The Royal Statistical Society, founded in 1834, together with many other statistical groups, was originally set up to just gather and publish data, as an attempt to reduce such uncertainty. It did not take long before statisticians wondered how the data might best be used and modern *statistical inference* was born. Charles Babbage, Adolphe Quetelet [...]. Among its famous members was Florence Nightingale, the society's first female member in 1858, whose work was shaped by this same intellectual climate. [...] Other notable RSS presidents have included William Beveridge, Ronald Fisher, which we will discuss 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 a clear distinction between the population under study and the selected sample, we face a problem. Neither the population—referred as the *true*—mean value, sometimes written as μ, nor its variance–referred as the *true* variance, and written as σ^2 are available to us. As we just saw, the *only thing we have is the finite set of observations in our sample*, hence we could try to build some "informative quantities" out of out data that would give us a hint of the central value, a measure of spread, etc. Such quantities are called *statistical estimators.* Common examples of such estimators are the *sample mean*, the *median*, and the *variance*, among others.

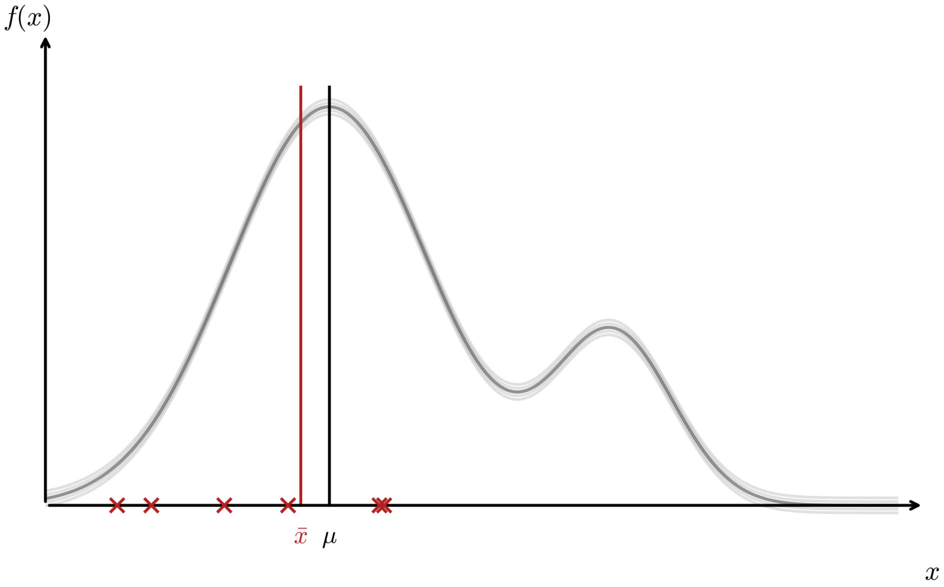


Figure 1.2: Representation of the *true* population mean μ, in black, and the observed *sample* mean $\bar{x}$. The true mean is and ideal and unaccessible quantity, while the sample mean can be computed as an estimator of the finite sample.

Once observations have been collected, a natural question arises: what is the *center*, or *typical* value of this data set? Mathematical quantities that measure the central tendency will be useful to summarize our data with a single representative number, providing an immediate sense of location within the distribution.

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 additional, more robust measures of central tendency.

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 meadian, given a size $n = 7$ would just be the midde (4th) value

$$M = 3 \ .$$

For instance, the data represented in LHS of Figure [...] 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.

A straightforward measure ofter used is the *mode*, the value—or values—that appear most frequently in the observation set. For the first sample $\mathcal{S} = \{1, 2, 3, 5, 3, 2, 7\}$ we just count the frequency of each value, and conclude that since both 2 and 3 occur most frequently, the dataset is *bimodal*, with modes 2 and 3. In the case of categorical data, such as eye colour or smartphone brands, the mode corresponds to the most common category.

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

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

The $n - 1$ in the denominator of (1.6) is called the Bessel correction factor, which ensures that only out of at leas $n = 2$ elements we can compute a finite variance. A more techical explanation is that it ensures that s^2 is an *unbiased estimator*, which we will discuss in Chapter 3

Note that the variance is just a sum of differences, and squared just so that we obtain a positive value. It is a measure starting with the first element (x_1, for $i = 1$) and until the last one (x_N, for $i = N$), of how far is each element from the mean value. If all elements in our sample are very close to the mean, then the sum of differences will be a small number, and we would get a variance s^2 close to zero. Meanwhile, if the elements are very different, we would obtain a larger variance.

Again, let's illustrate with an example. If we compute the variance of our very first example set $\mathcal{S} = \{1, 2, 3\}$, which has just $n = 3$ observations, we get

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

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.

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

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

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

Finally, *quantiles* divide the ordered data into equal proportions. The p-th quantile Q_p is the value below which a fraction p of the data lies. Special cases include the *first quartile* (Q_1, 25th percentile), the *median* (Q_2, 50th percentile), and the *third quartile* (Q_3, 75th percentile). A rigurous definition of quantiles requires the idea of distribution and cumulative probability, so we will discuss them in next chapter. As a note, 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, mean, median, mode, variance, standard deviation, and quantiles provide a rich, complementary view of the dataset's central tendency and variability, allowing for both numerical and graphical summaries that capture the essence of the data.

Variation is not merely a technicality; it is the very essence of uncertainty. Without spread, probability would be trivial, for every outcome would be the same. It is in the differences among observations that statistical inquiry finds its substance. Hence, central tendency and variation together provide the complementary lenses through which data becomes intelligible. They allow us to say whether two groups are alike or unlike, whether a new result is ordinary or surprising, whether the observed variation is too great to be dismissed as chance. In this sense, descriptive statistics foreshadows the inferential methods to come, hinting at deeper laws beneath the numbers.

1.3 Data visualization

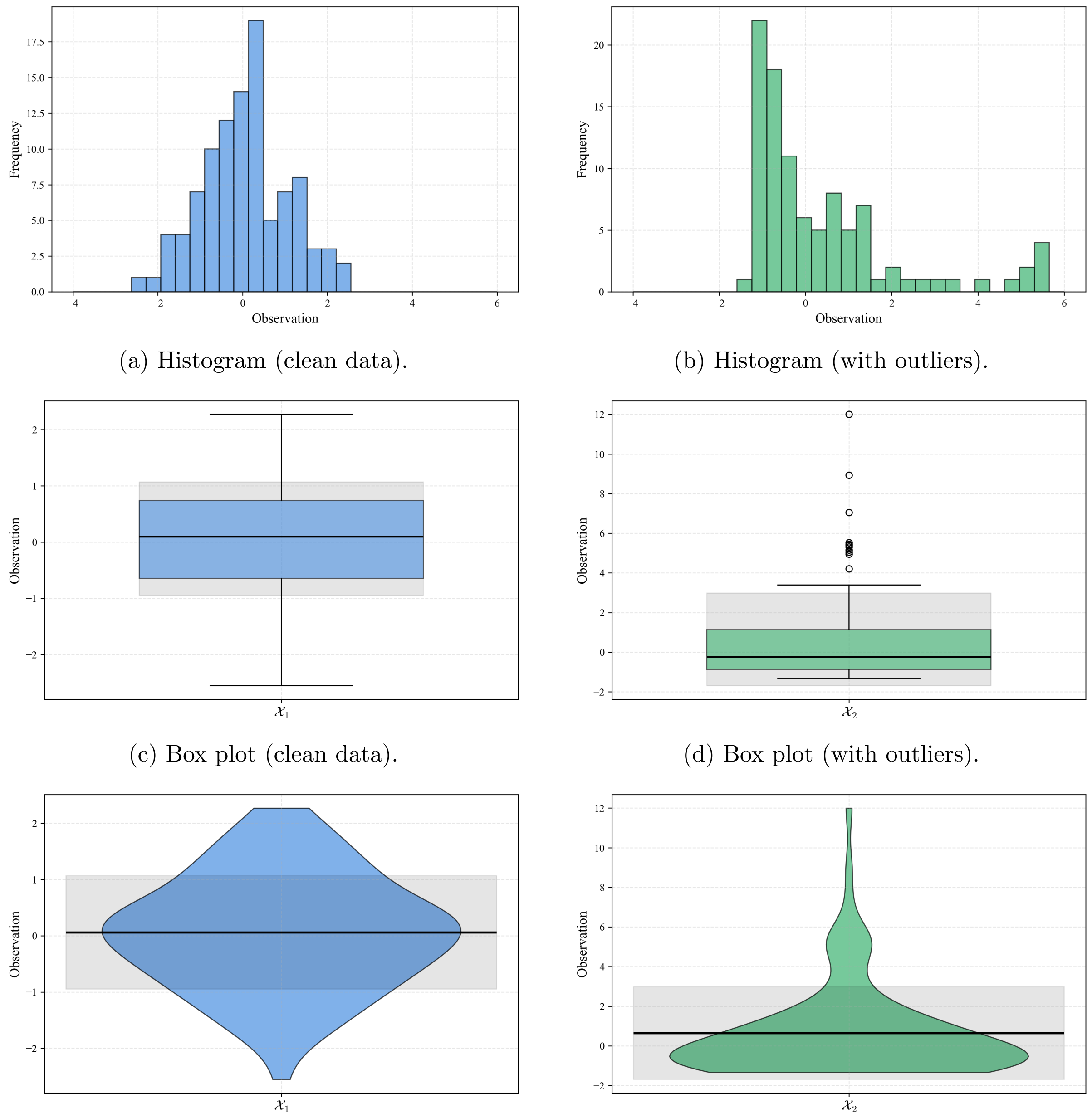
While numerical summaries are useful, the human mind often understands patterns much faster through vision than calculation. By *data visualization* we mean a series of techniques used to transform numbers and sequences into shapes, colours and structures that are easier to interpret, and that can be grasped at a glance. It turns abstraction into perception and often reveals regularities invisible to formulas alone. Nowadays, a broad series of fields falling under the name of data visualization - or data *representation* - have become among the pillars of any scientific or data related topic.

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].



(a) Histogram (clean data).

(b) Histogram (with outliers).

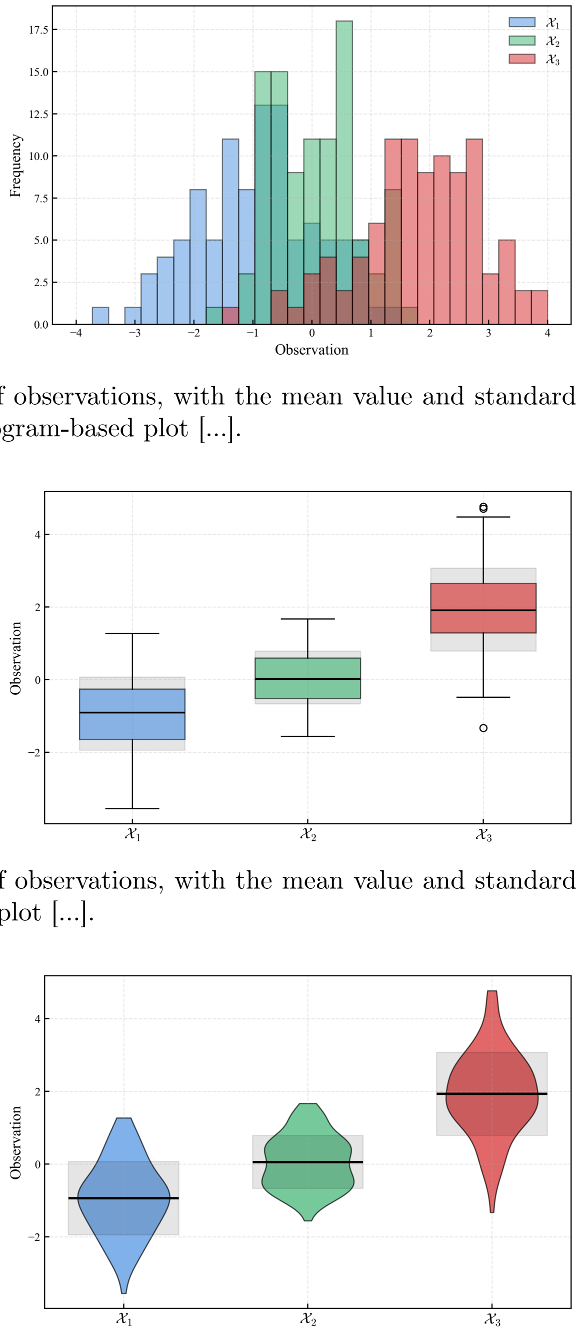
(c) Box plot (clean data).

(d) Box plot (with outliers).

(e) Violin plot (clean data).

(f) Violin plot (with outliers).

Figure 1.3: 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.



(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 1.4: 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 seventeenth 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. Subsequent refinements by De Moivre and Laplace transformed probability into a powerful analytical theory, while its fully axiomatic structure only crystallised in the twentieth century, as we will see.

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. 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.

A decisive step toward mathematical rigor was taken by Andrey Kolmogorov in 1933. In his *Grundbegriffe der Wahrscheinlichkeitsrechnung* [15], Kolmogorov showed that probability could be treated as a branch of measure theory, independent of any specific interpretation. Rather than defining probability by intuition, symmetry, or frequency, he postulated a small set of axioms from which the entire formal theory follows.

This axiomatic approach provided a common mathematical framework within which classical, frequentist, and Bayesian interpretations could coexist. While philosophical disagreements about the meaning of probability persist, Kolmogorov's formulation ensures that all interpretations obey the same internal rules of consistency. In this sense, modern probability theory is less concerned with what probability *means* and more with how probabilistic reasoning must behave if it is to be logically coherent.

The explicit mathematical formalization of decision-making under uncertainty is, however, a relatively recent development. It is usually attributed to the British mathematician Frank P. Ramsey (1903–1930), who in his 1926 paper *Truth and Probability* [16] introduced a subjective interpretation of probability grounded in rational preference. Ramsey showed that coherent choices imply numerical probabilities and utilities, thereby laying the foundations of expected utility theory. His work marked a shift from viewing probability solely as a property of random mechanisms to treating it as a rational measure of belief.

Parallel developments in the early twentieth century—most notably by Pearson, Fisher, and

Neyman—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, 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 [15]. A probability $\mathbb{P}$ assigns a number to each event and satisfies 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 $\{x_1, x_2, \dots\}$, these rules imply the normalization condition

$$\sum_i \mathbb{P}(x_i) = 1,$$

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 [17]. 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 Bayes [18] and developed further by Laplace and later authors such as de Finetti [19] and Jaynes [20]. 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 both rely on Kolmogorov’s axioms. The difference lies not in the calculations, but in how probability statements are interpreted. Bayesian methods and their practical consequences will be introduced formally in later chapters.

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

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

2.2.1 Bernoulli trials

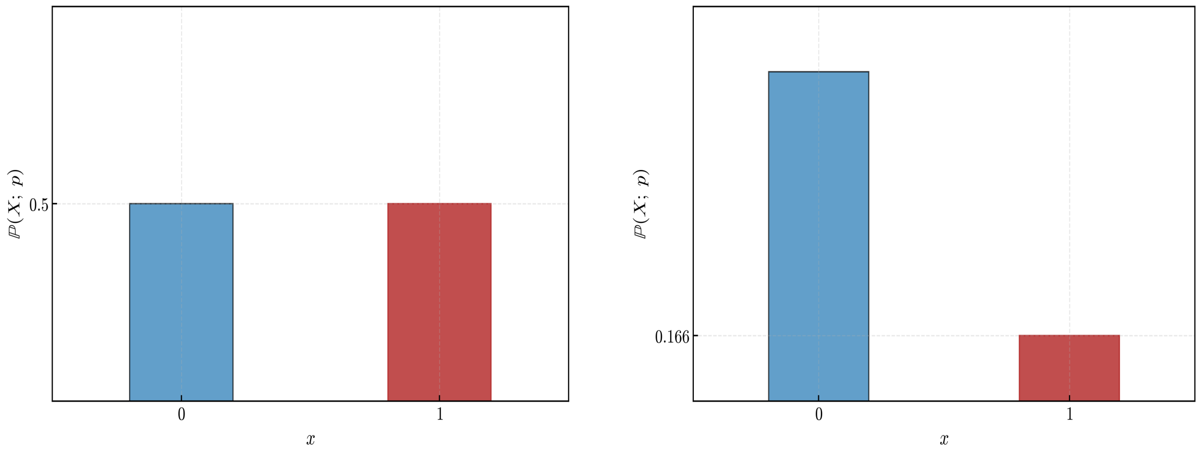
The Bernoulli trial was formalized by Jacob Bernoulli in *Ars Conjectandi* (1713) [21]. His motivation was to understand how regularity emerges from randomness when an experiment with two outcomes is repeated many times.

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 $p = 1/2$.



(a) Bernoulli example 1.

(b) Bernoulli example 2.

Figure 2.1: Probability distribuion

2.2.2 Discrete uniform distribution

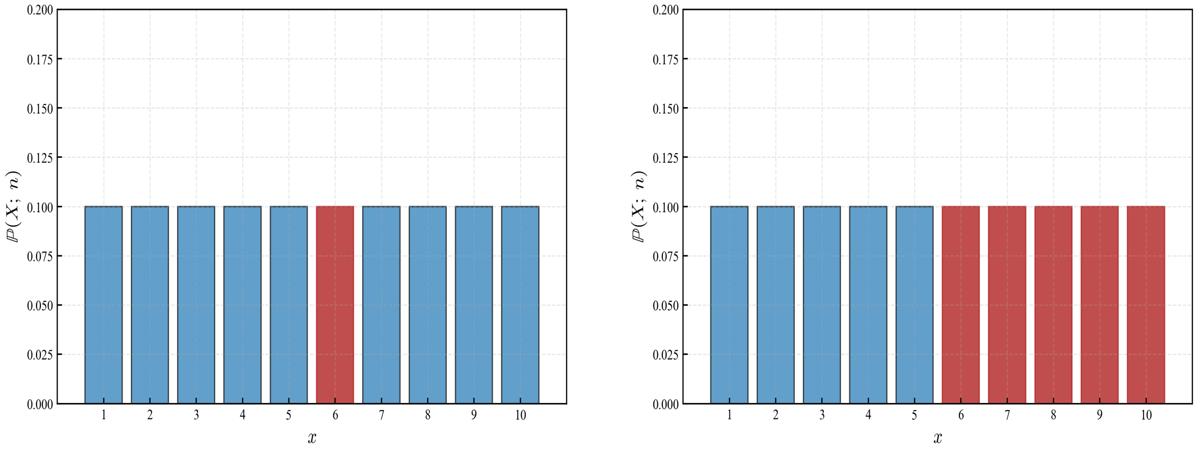
The discrete uniform distribution has its roots in classical symmetry arguments used in early probability theory. It formalizes the idea that, in the absence of distinguishing information, all outcomes should be treated equally. This idea reflects a principle already present in early probability theory: 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 $\{x_1, \dots, x_n\}$, the discrete uniform distribution assigns

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

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 distribiuon

2.2.3 Binomial distribution

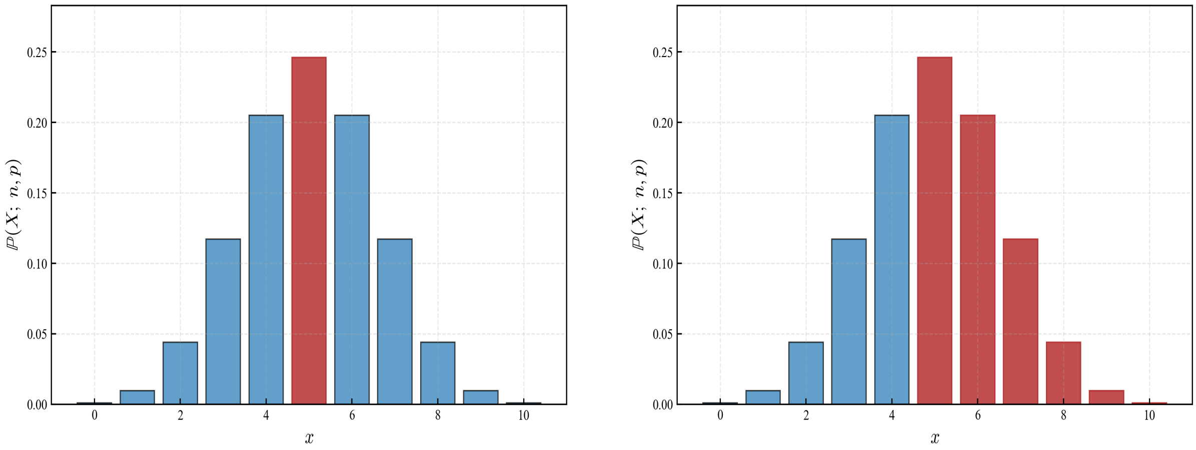
The binomial distribution was systematically studied by Abraham de Moivre in the early eighteenth century. His analysis of repeated Bernoulli trials led not only to the binomial formula but also to the first appearance of the normal approximation. De Moivre introduced the binomial distribution while 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

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

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 distribuion

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

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

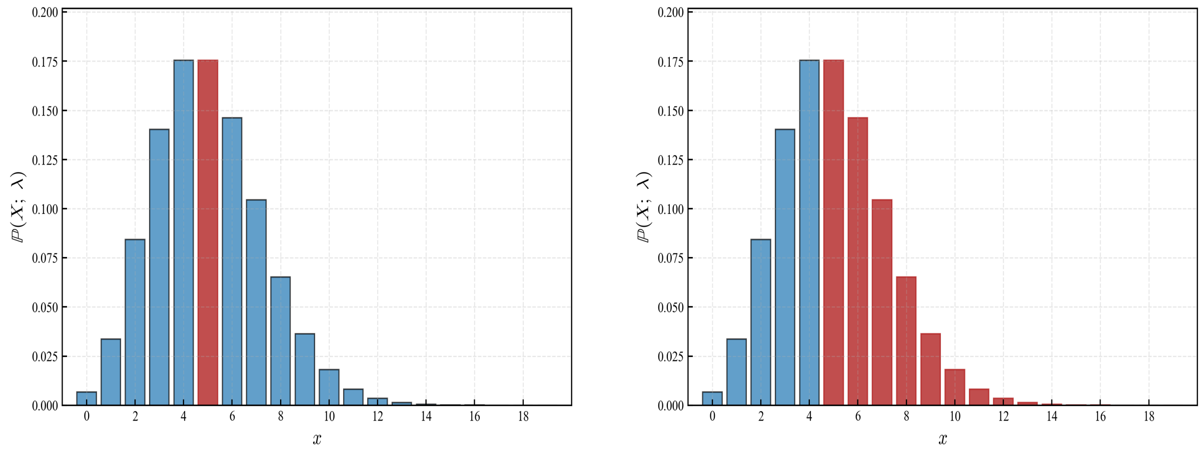
where $\lambda > 0$ is the average rate of occurrence.

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.

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.



(a) Probability distribution.

(b) Cumulative probability.

Figure 2.4: Probability distribuion

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

$$f(x) \geq 0, \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1. \tag{2.6}$$

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 eighteenth 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) [22], 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.

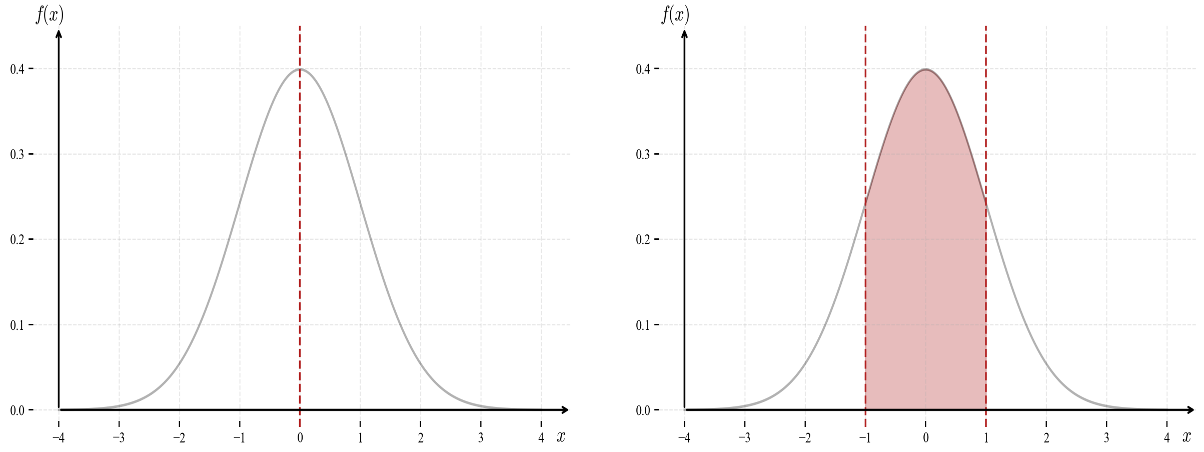
The Gaussian distribution models the accumulation of many small, independent effects. Its central role in probability theory is explained by the central limit theorem, which establishes it as a universal limiting distribution.

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

$$f(x;\ \mu,\ \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}. \tag{2.7}$$

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 distribiuion

2.3.2 Exponential distribution

The exponential distribution arose in the nineteenth 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 [23, 24]. 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.

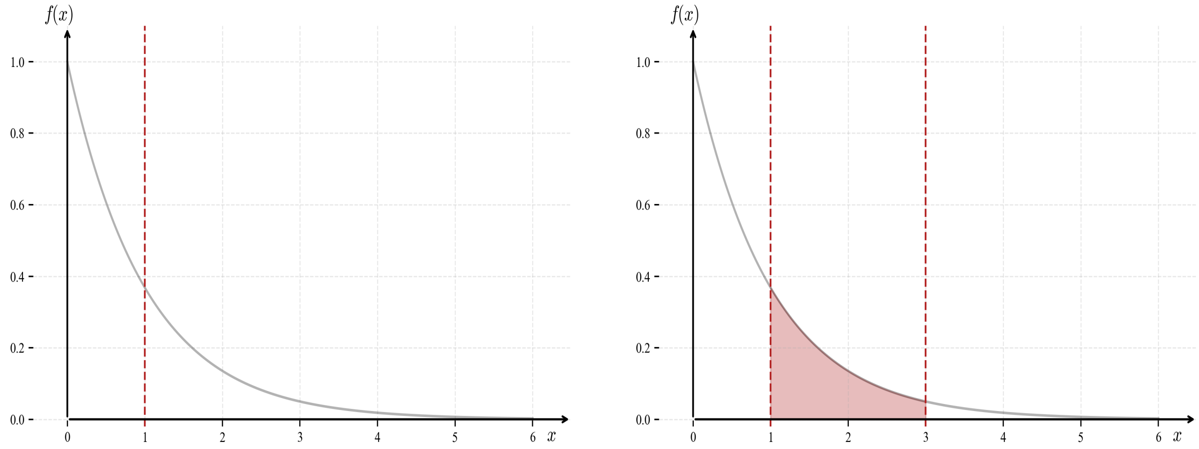
It naturally models the time until the first occurrence of a random event and is characterized by the absence of memory: the future waiting time does not depend on how much time has already elapsed.

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

$$f(x;\ \lambda) = \lambda e^{-\lambda x}, \qquad x \geq 0. \tag{2.8}$$

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.



(a) Probability distribution.

(b) Cumulative probability.

Figure 2.6: Probability distribuion

2.3.3 Continuous uniform distribution

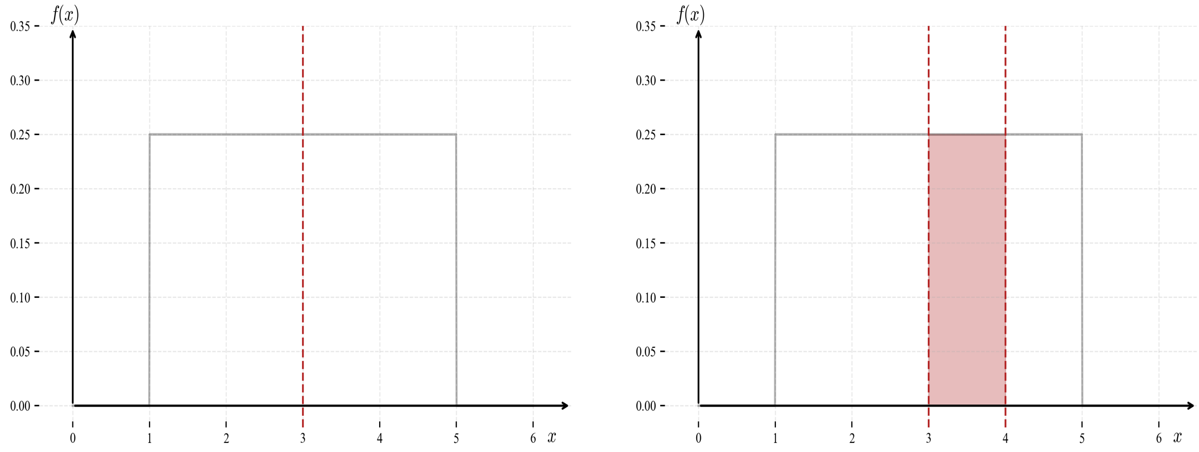
The continuous uniform distribution extends classical symmetry arguments already present in Laplace's *Théorie Analytique des Probabilités* (1812) [25]. It represents complete ignorance about where within a bounded interval an outcome may fall. The continuous uniform distribution formalizes the idea 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

$$f(x;\ b,\ a) = \frac{1}{b-a}, \qquad a \leq x \leq b. \tag{2.9}$$

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.

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 distribuion

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.

The most fundamental moment is the *expected value*, also called the 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

$$\mathbb{E}[X] = \sum_i x_i \, \mathbb{P}(X = x_i). \tag{2.10}$$

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

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \, f(x) \, dx. \tag{2.11}$$

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,

$$\mathbb{E}\left[\sum_i X_i\right] = \sum_i \mathbb{E}[X_i], \tag{2.12}$$

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:

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]. \tag{2.13}$$

An equivalent and often more convenient expression for the variance is

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2. \tag{2.14}$$

For a discrete random variable, this corresponds to

$$\mathrm{Var}(X) = \sum_i (x_i - \mu)^2 \, \mathbb{P}(X = x_i), \qquad \mu = \mathbb{E}[X], \tag{2.15}$$

while for a continuous random variable it is given by

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \, f(x) \, dx. \tag{2.16}$$

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

$$\mathbb{E}[X] = p, \tag{2.17}$$

$$\mathrm{Var}(X) = p(1-p). \tag{2.18}$$

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

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

$$\mathbb{E}[X] = np, \tag{2.19}$$

$$\mathrm{Var}(X) = np(1-p). \tag{2.20}$$

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

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

$$\mathbb{E}[X] = \lambda, \tag{2.21}$$

$$\mathrm{Var}(X) = \lambda. \tag{2.22}$$

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

$$\mathbb{E}[X] = \frac{n+1}{2}, \tag{2.23}$$

$$\mathrm{Var}(X) = \frac{n^2-1}{12}. \tag{2.24}$$

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

$$\mathbb{E}[X] = \mu, \tag{2.25}$$

$$\mathrm{Var}(X) = \sigma^2. \tag{2.26}$$

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

$$\mathbb{E}[X] = \frac{1}{\lambda}, \tag{2.27}$$

$$\mathrm{Var}(X) = \frac{1}{\lambda^2}. \tag{2.28}$$

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

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

$$\mathbb{E}[X] = \frac{a+b}{2}, \tag{2.29}$$

$$\text{Var}(X) = \frac{(b-a)^2}{12}. \tag{2.30}$$

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.

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Chapter 3

Estimation, variability and confidence

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.* Ancient Sources Edition, 45 BCE.

[10] Gerolamo Cardano. *Liber de Ludo Aleae.* Apud Joannem Baptistam Ferrarium, Paris, 1663.

[11] Keith Devlin. *The Unfinished Game.* Basic Books, 2008.

[12] Christiaan Huygens. *De Ratiociniis in Ludo Aleae.* Elzevier, Leiden, 1657.

[13] Jacob Bernoulli. *Ars Conjectandi.* Thurneysen Brothers, Basel, 1713.

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

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

[16] Frank P. Ramsey. Truth and probability. In D. H. Mellor, editor, *The Foundations of Mathematics and Other Logical Essays*, pages 156–198. Routledge and Kegan Paul, London, 1926.

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

[18] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 1763.

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

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

[21] Jacob Bernoulli. *Ars Conjectandi.* Thurneysen Brothers, Basel, 1713.

[22] Carl Friedrich Gauss. *Theoria Motus Corporum Coelestium.* 1809.

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

[24] Joseph L. Doob. *Stochastic Processes.* Wiley, 1953.

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