Imperial College London

Introduction to probability theory and statistical inference

Jesús Urtasun Elizari

Research Computing and Data Science

September 2, 2025

Contents

Introduction

The theory of probabilities is at bottom nothing but common sense reduced to calculation.

— Pierre-Simon Laplace

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 data analysis to modelling in physics, biology, machine learning and quantum mechanics, among many others.

As mentioned, the course is organised in five chapters. In the first two we will introduce the idea of sampling, probability and random events with simple and intuitive examples, and we will see how different approaches have been used to model information and chance in different times. The introduction here is twofold. Chapter 1 *Descriptive statistics* introduces the idea of uncertainty, sampling, and central tendency, aiming to describe and understand populations and observation sets, while Chapter 2 *Probability and random events* focuses on the mathematical definition of probability. Here we will cover the idea of random processes - also referred to as *stochastic*, probability and distribution, as a set of tools that enables mathematical predictions in uncertain cases, such as the *expected value*.

Chapter 3 *Parameter estimation* will introduce the essential difference between prediction and inference, and revisit the concept of sampling and population in more detail. We will discuss how to build *estimator* quantities out of our samples, as a way to reconstruct - or *infer* - the underlying phenomena of a population given a finite set of observations. Here we will see some general results which may sound familiar already, such as *"The Law of Large Numbers"*, the *"The Central Limit Theorem"*, or the *"Maximum Likelihood Estimation"* method.

In Chapter 4 we will discuss a group of topics commonly referred to as *hypothesis testing*. Here we will introduce the idea of hypothesis, how to quantify certainty and bias, how to model significance with the so-called *p-values*, and some common examples of statistic tests. Chapter 5 will cover with some detail conditional and Bayesian probability, revisit the idea of stochasticity and introduce the so called Markov processes.

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].

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 value, 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"*), and Jacob Bernoulli, whose 1713 *"Ars Conjectandi"* (*"The Art of Conjecturing"*) remains among the most influential early texts in the field. Their works, alongside with many others, collectively laid the groundwork for the probabilistic and statistical methods that foreshadow modern scientific reasoning [12], [13], [14].

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

A large part of history of science could be summarized as an effort to translate observations of reality into precise, mathematical understanding. A record of the continuous human striving for a formulation and description of the real world in mathematical terms. To define mathematically the phenomena we find in the natural world, 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, to estimate the number of stars in the observable universe, relating the boiling point of water to the external pressure, or the number of lung cancer patients to pollution levels around smoking areas.

Colombian mathematician Luis C. Recalde marvellously summarizes the mathematical endeavour as three core tasks. For him, mathematics could be reduced to all tasks related to count, measure, and sort. When it comes to the description of populations, sampling, and chance, the fields of statistics and probability develop ideas such as randomness, relationship, correlation, confidence and reproducibility, among others. Inspired by Recalde's aim to simplify, we could summarize all statistical issues as concern with *uncertainty*, or *variation* among observations.

Hence, a philosophical position often adopted is that statistics is essentially the study of uncertainty, and that the statistician's role is to assist workers in 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.

Historically, uncertainty has been associated with games of chance and gambling. The Royal Statistical Society, together with many other statistical groups, was originally set up to gather and publish data, as an attempt to reduction in uncertainty. It remains an essential part of statistical activity today and most Governments have statistical offices whose function is the acquisition and presentation of statistics. It did not take long before statisticians wondered how the data might best be used and modern *statistical inference* was born.

The mathematical formalization of decision-making is actually quite a recent development. It is usually attributed to British mathematician Frank P. Ramsey (1903–1930), who in his 1926 paper *Truth and Probability* [16] introduced a formal, subjective interpretation of probability, laying the groundwork for what later became expected utility theory in decision-making under uncertainty. In short, Ramsey formalized how rational agents should assign probabilities and make decisions based on personal beliefs and preferences. All starting from the apparantley-simple question *'how should we make decisions in the face of uncertainty?'*.

1.1 Sampling and data types

All statistical inquiries begins 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* $\mathcal{P}$, and the *sample* $\mathcal{S}$. By *population* we mean the complete set of all possible observations under study, normally written as

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

The *sample*, 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 normally denoted by the upper case N - is defined as

$$\mathcal{S} = \{x_{i_1}, x_{i_2}, \ldots, x_{i_n}\}, \quad n < N \, , \tag{1.2}$$

where the i-subscripts remind us that the sample consists of selected observations from the population, not necessarily consecutive or all of them. The population represents the ideal object of inference, while the sample is the concrete, finite evidence available to us. 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 are of the same kind. 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 group is normally referred to as numerical - or *quantitative* - data. These measure numerical quantities and are 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.

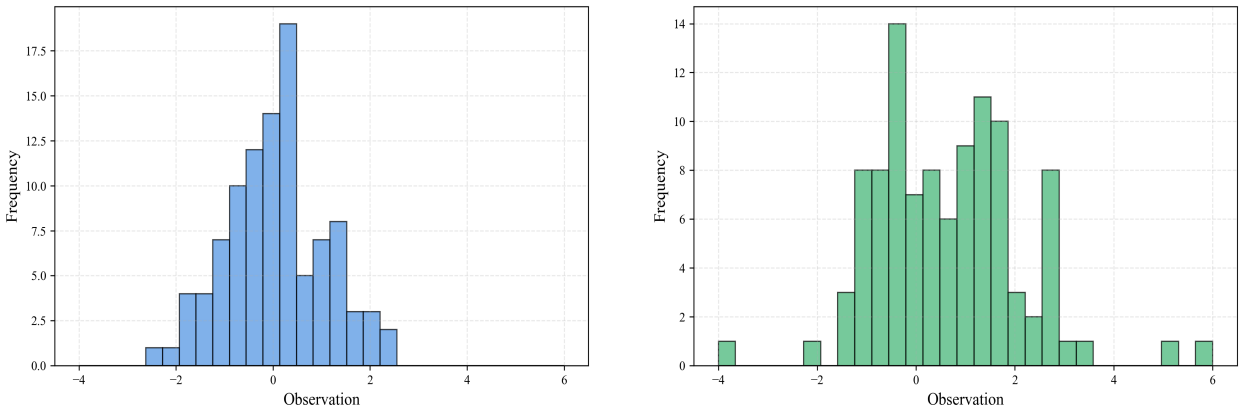
Distinguishing between these types is no mere slang; different types of observations require different mathematical tools, and will be described in different ways. For example, it would not make sense to compute a mean out of smartphone brands, but to compute the mean of their prices is informative. 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.

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 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.



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

Figure 1.1: Histogram representation of $n = 100$ observations drawn from a Gaussian distribution. The RHS shows a clean Gaussian distribution, symmetric around the central value and with no outlier points, while the LHS shows a skewed version, where some outliers make the distribution deviate from the symmetric case.

The *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 number of tomatoes in a bag, 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, or our *sample* $\mathbf{x}$. We could simply write it as a list - or a *vector* - in the following way:

$$\mathbf{x} = \{x_1, x_2, \ldots, x_n\} \, .$$

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 a quantity called the *mean* - or *average* - 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:

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

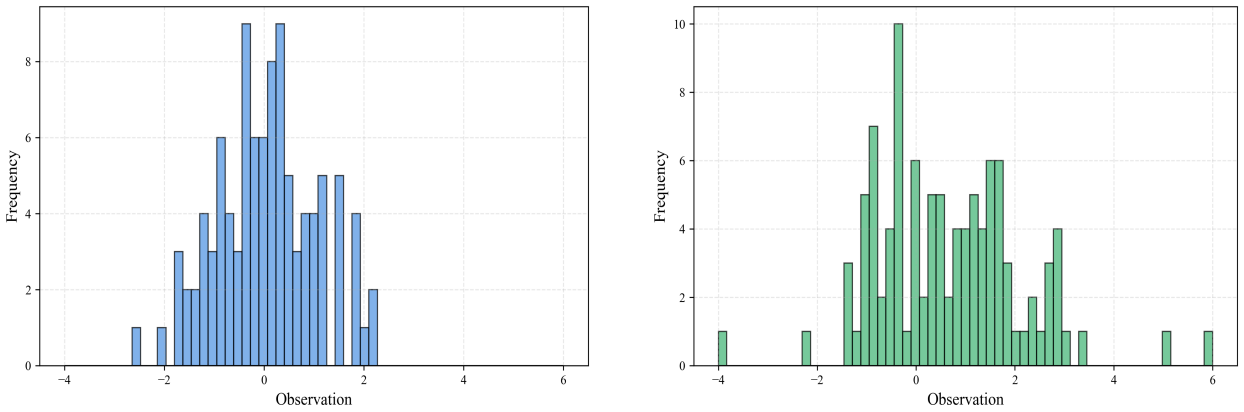
We can write this in a slightly more compact way as a *summation*, as follows:

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

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 pause here for a second, and give a note about notation. Remember the difference we made at the very beginning between sample and population, as notations may differ between different books and literature sources. Normally, the sample mean is written just as (1.4), while for the full population of N elements $x_1, x_2, \ldots, x_N$ - before any sampling - the *population mean* is normally denoted as μ, and defined accordingly

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \, . \tag{1.5}$$



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

Figure 1.2: Histogram representation of $n = 100$ observations drawn from a Gaussian distribution. The RHS shows a clean Gaussian distribution, symmetric around the central value and with no outlier points, while the LHS shows a skewed version, where some outliers make the distribution deviate from the symmetric case. Narrower binning leads to higher resolution, but it is more sensitive to outliers.

We will see more about the difference between sample mean and population mean when we discuss parameter estimation in Chapter 3. For now just keep in mind that $\bar{x}$ is the mean of our sample of just n drawn observations, while μ refers to the mean of the idealized, complete population.

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 $\mathbf{x} = \{1, 2, 3\}$, 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 $\mathbf{x} = \{4, 5, 6\}$. Substituting into the general expression (1.4) gives

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

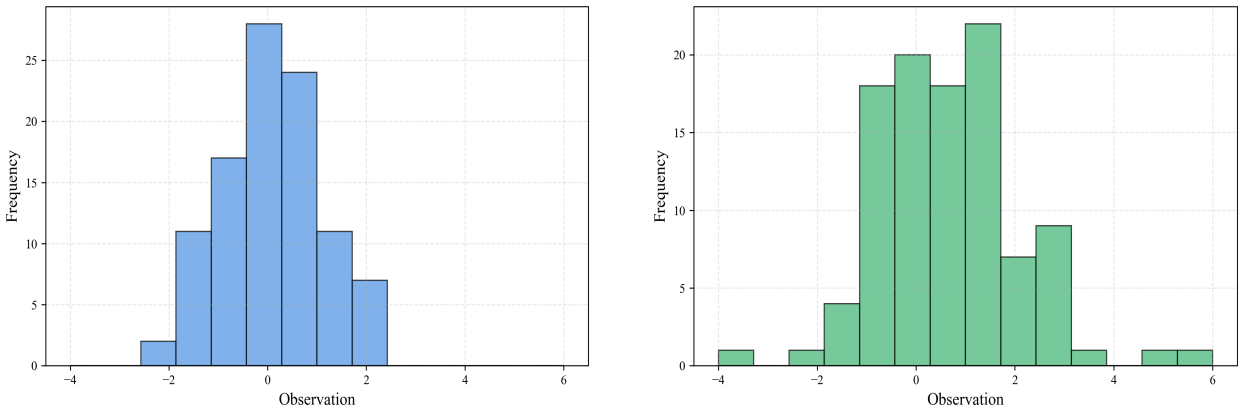
The mean captures information about the "central" value, where most events cluster. Although useful, it is sensitive to extreme values or *outliers*, which motivates the definition additional, more robust measures of central tendency.

The *median* represent 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

$$M = \begin{cases} x_{(k+1)} \, , & \text{if } n = 2k + 1 \text{ (odd)} \, , \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} \, , & \text{if } n = 2k \text{ (even)} \, . \end{cases} \tag{1.6}$$

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

$$\mathbf{x}_{\text{ordered}} = \{1, 2, 2, 3, 3, 5, 7\} \, . \tag{1.7}$$



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

Figure 1.3: Histogram representation of $n = 100$ observations drawn from a Gaussian distribution. The RHS shows a clean Gaussian distribution, symmetric around the central value and with no outlier points, while the LHS shows a skewed version, where some outliers make the distribution deviate from the symmetric case. Thicker binning usually implies smaller resolution, but averages the raw observations and it remains more robust against outliers.

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

$$M = x_{(4)} = 3 \, . \tag{1.8}$$

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

$$\mathbf{x}_{\text{ordered}} = \{1, 2, 2, 3, 3, 4, 5, 7\}. \tag{1.9}$$

With has an even number of elements now, $n = 8$. Hence, applying such case in (1.6), the median is the average of the two middle values

$$M = \frac{x_{(4)} + x_{(5)}}{2} = \frac{3+3}{2} = 3 \, . \tag{1.10}$$

Unlike the mean, the median is robust to outliers and skewed data, capturing the central position of the dataset even with repeated values. For instance, the data represented in LHS of Figure 1.1 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 *mode* is the value - or values - that appear most frequently in the observation set, which is quite a straightforward measure. For the first sample $\mathbf{x} = \{1, 2, 3, 5, 3, 2, 7\}$ 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 *variance* s^2 of a set as a quantity that captures how far are the elements from the mean value,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \, , . \tag{1.11}$$

and again, we will use a different notation for the *population variance*

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \, . \tag{1.12}$$

If we pay close attention, we see that the definitions of s^2 and σ^2 are not identical. The $n-1$ in the denominator of (1.11) is called the Bessel correction factor, and it arises from the fact that treating finite samples is not the same as referring to the complete population. We will return to this topic in Chapter 3, when we discuss the concept of estimators and Maximum Likelihood Estimation.

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 $\mathbf{x} = \{1, 2, 3\}$, 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 $\mathbf{x} = \{4, 5, 6\}$. By substituting in the general expression (1.11) 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 $\mathbf{x}$ 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,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ , \tag{1.13}$$

and for the entire population,

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}\ . \tag{1.14}$$

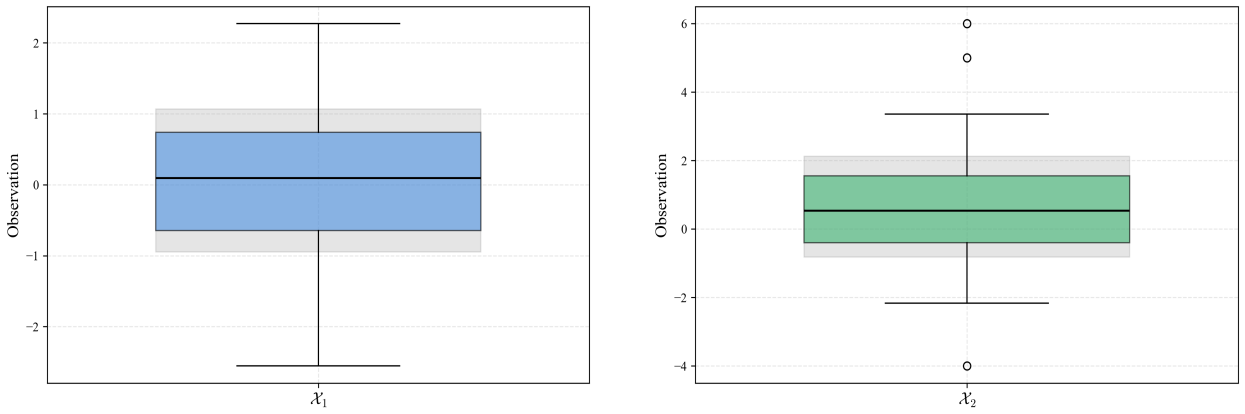
At a glance, variance and standard deviation quantify how much the elements of a dataset deviate from the mean, capturing the notion of *spread*.

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). Formally, for a continuous cumulative distribution function (CDF) F, the p-th quantile satisfies

$$Q_p = \inf\{x : F(x) \geq p\}. \tag{1.15}$$

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.



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

Figure 1.4: Box plots representing $n = 100$ observations drawn from a Gaussian distribution. The central black line shows the mean value, representing the central tendency where the bulk of events lie. The shadowed area highlights the standard deviation, as measure of the variability and spread the observations with respect to the mean

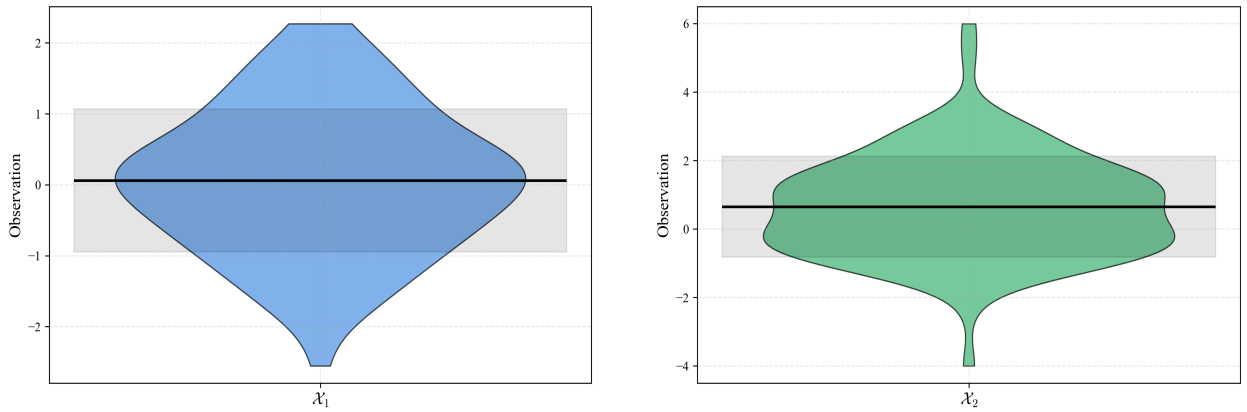
1.3 Data visualization

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.

The *histogram* is found among the oldest and most fundamental visualization tools. The concept of dividing data into intervals to visualize frequency dates back to Karl Pearson in the late 19th century, who formalized it as a graphical representation of probability distributions [17]. A histogram divides the range of a dataset into consecutive intervals, or *bins*, and represents the the amount - or relative *frequency*- of observations falling within each bin as the height of a bar. This simple yet powerful plot provides an immediate visual impression of the dataset's distribution, allowing one to identify symmetry, skewness, concentration of values, and potential gaps. For example, a symmetric histogram, like the one in LHS of Figure 1.1 suggests a roughly balanced distribution around the mean, while a right-skewed histogram, like the one displayed in the RHS of Figure 1.1, indicates that higher values are less frequent - or less *probable* - but can yet influence measures like the mean. This is the state of the art in physical sciences and whenever data is supposed to fit a mathematical prediction.

Building a histogram in an informative way is extremely powerful, and there are some subtleties to consider. As a rule of thumb, look for natural divisions in the data, and keep all bins the same size, covering the whole range under study. Outliers can skew, so they must be treated carefully. Figures 1.2 and 1.3 show how the binning size can affect the distribution of data. Smaller binning leads to more resolution but can be easily distorted in the presence of outliers, while few large bins are robust agains though losing the accuracy in resolution. For skewed distributions it is normally better to use the median and the IQR.

The box plot, also known as the *box-and-whisker* plot, was introduced by John Tukey in 1970 as part of his work on exploratory data analysis [18]. The box plot offers a compact summary of a dataset's central tendency, spread, and potential outliers. Constructed from five key statistics - the minimum, first quartile (Q_1), median (Q_2), third quartile (Q_3), and maximum - it clearly shows the *interquartile range* (IQR $= Q_3 - Q_1$) and highlights points that fall outside 1.5 times the IQR as outliers. This representation allows for quick comparisons across multiple groups, and it is particularly useful for detecting asymmetry and variability without being overly influenced by extreme values. It is widely used in biological and clinical



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

Figure 1.5: Violin plots representing $n = 100$ observations drawn from a Gaussian distribution. The central black line shows the mean value, representing the central tendency where the bulk of events lie. The shadowed area highlights the standard deviation, as measure of the variability and spread the observations with respect to the mean

sciences where an experiment can be repeated many times with relatively small sizes. Figure 1.4 displays the same data represented as histograms in Figures 1.1 as box plots.

Finally, the *violin* plot is a more recent innovation, combining the box plot with a series of mathematical tools that represent as well the shape of the distribution. While the precise origin is less formally documented, it gained prominence in the late 20th century in statistical software environments, such as R, during the 1990s [19]. Essentially, the violin plot extends the concept of the box plot by combining it with a kernel density estimate of the data. This plot not only displays the median and quartiles but also provides a smooth depiction of the distribution's shape, revealing features such as multimodality or skewness that might be obscured in a simple box plot. By showing both summary statistics and the underlying density, the violin plot gives a richer, more nuanced view of the dataset, particularly when comparing several groups side by side. As an example of such comparison, see Figure [...].

Underlying these graphics we have mentioned the concept of *quantiles*. Put simple, quantiles provide a language for describing position of an observation within a distribution. Mathematically, the p-quantile of a dataset is the value q p such that at least a proportion p of the data lies below it. The median is the 50th percentile, quartiles mark the 25th and 75th percentiles, and finer partitions yield deciles or percentiles.

To conclude, let us emphasize that graphs and diagrams are not mere decoration but deeply useful instruments of analysis. They allow patterns to leap from obscurity in shallow data, invite hypotheses, and sometimes contradict assumptions and expectation. In practice, visualization is both a beginning and a test: a first impression of data, and a final check on the reasonableness of results derived through calculation.

1.4 Dependency, linearity, correlation

Data rarely lives in isolation. Often, even in the simplest case, one variable depends upon another. Rainfall influences crop yields, study hours affect exam results, in the same way the brightness of a star relates to its temperature, and atmospheric carbon levels affect global temperatures, just to list some examples. Hence, recognizing and describing such dependencies lies at the heart of descriptive statistics and prepares the way for predictive models.

The simplest and most widely studied form of dependency is *linearity*. When one variable y tends to increase in proportion to another x, the relation can be sketched as a straight line. In the language of calculus - also called some times *analysis*, or *regression* - the best-fitting line is expressed as

$$y(x) = ax + b \,, \tag{1.16}$$

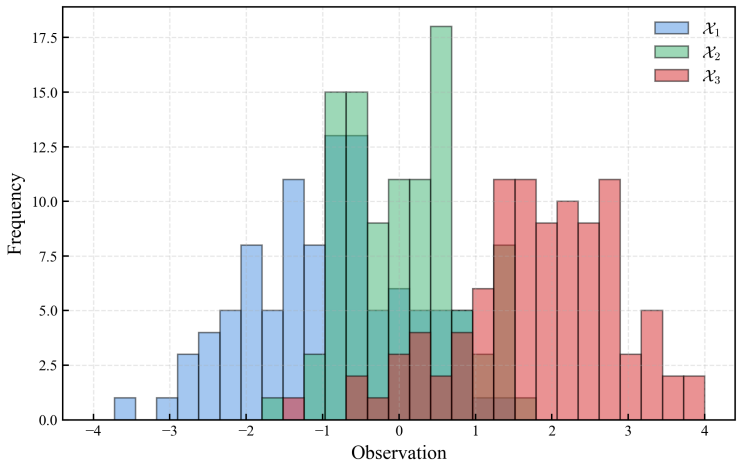


Figure 1.6: Three sets of observations, with the mean value and standard deviation represented as a violin plot [...]. The red line shows the mean value, representing the central tendency where the bulk of events lie. The shadowed area highlights the standard deviation, as measure of the variability and spread the observations with respect to the mean. In this case the three observation sets, or *samples*, consist of 100 observations. In the case of χ_1, χ_2, χ_3 [...].

where a is called the *slope* and b stands for the *intercept*, or *independent term*. Different literature sources and fields of study name these variables slightly different, so it is worth to pause here for a second. Mathematicians and physicists use to call x simply the *independent* variable, and y the *dependent* one, often writing it as $y(x)$. Meanwhile, biologists and clinical scientists often use the term *explanatory* variable for x, and *response* variable for y, which can feel a bit odd at first. The reason for such naming is that, in the same way physicists would say that $y(x)$ *depends* on x, clinicians would say that it is a *response*, or *explained by* x. Figure [...].

Yet, not all relationships in nature are as simple as the linear. Indeed, many processes curve and deviate from a straight dependency. For instance, the trajectory of a projectile follows a *quadratic* path, as does the kinetic energy of a particle with respect to its velocity, or the area of a square, which depends quadratically on the length of its side. The quadratic dependency can be written mathematically as

$$y(x) = ax^2 + bx + c \,, \tag{1.17}$$

where the squared term introduces the curvature. More generally, one may consider *polynomial* relationships, where higher powers of x capture increasingly intricate bends in the data

$$y(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \,. \tag{1.18}$$

Beyond polynomials, mathematical modelling admits much broader and richer forms, such as exponential growth and decay, logarithmic compression, trigonometric oscillations, and nonlinear interactions of multiple variables. Such models can are used describe the subtleties of physical, biological, and social systems with remarkable fidelity. Yet, statistics has long placed linearity at its center, for two main reasons: mathematical convenience and interpretive clarity. Straight lines are tractable - they allow for analytic and simple solutions, and clear geometric intuition. More importantly, linear models often suffice as approximations, capturing the dominant trend even when the world curves beneath. For these reasons, linearity remains the default language of statistical dependency, and the first lens through which we attempt to see structure in scattered data.

The idea of *correlation* is a fundamental measure of association between two random variables, quantifying how strongly they vary together. The most widely used mathematical description is the *Pearson correlation coefficient*, introduced by Karl Pearson in the 1890s, which is built upon the idea of covariance normalized

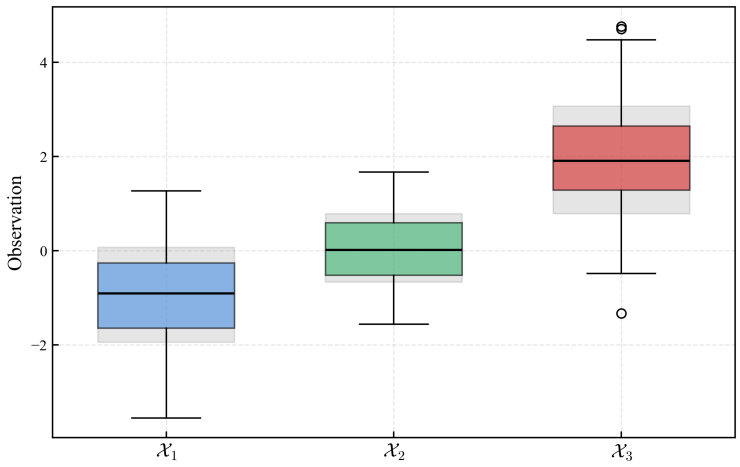


Figure 1.7: Three sets of observations, with the mean value and standard deviation represented as a box plot [...]. The red line shows the mean value, representing the central value where the bulk of events lie, and the shadowed area the standard deviation, as measure of the variability, or how spread the observations are with respect to the mean. In this case the three observation sets, or *samples*, consist of 100 observations. In the case of χ_1, χ_2, χ_3 [...].

by variability. For two random variables x and y, the population correlation $\rho_{x,y}$ is defined as

$$\rho_{x,y} = \frac{\mathrm{Cov}(x,y)}{\sigma_x\,\sigma_y} \tag{1.19}$$

where the covariance is defined as

$$\mathrm{Cov}(x,y) = \mathbb{E}[(x-\mu_x)(y-\mu_y)]\ . \tag{1.20}$$

Here $\mu_x = \mathbb{E}[x]$ and $\mu_y = \mathbb{E}[y]$ denoting the *expected values* of x and y. The expected value of a random variable is, intuitively, its long-term average or center of mass; it summarizes the *typical* value the variable takes. For a discrete random variable x with outcomes x_i and probabilities p_i, the expected value is

$$\mathbb{E}[x] = \sum_i p_i x_i\ . \tag{1.21}$$

The computation of expected values is not such a trivial topic. In next chapter we will revisit this concept with more detail, in the context of random variables and probability distributions. For now, and for practical purposes with finite datasets, we can approximate the expected value just by the *arithmetic mean*, or *sample mean* of the observations, $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\ , \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\ .$$

Let's illustrate with an example. Consider the dataset

$$x = \{1,2,3\}\ , \quad y = \{2,4,5\}\ . \tag{1.22}$$

The sample means are trivial to compute

$$\bar{x} = \frac{1+2+3}{3} = 2\ , \quad \bar{y} = \frac{2+4+5}{3} = 3.67\ ,$$

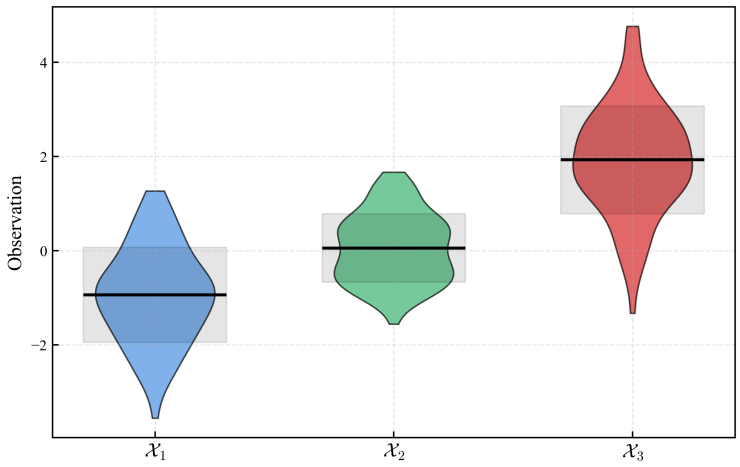


Figure 1.8: Three sets of observations, with the mean value and standard deviation represented as a violin plot [...]. The red line shows the mean value, representing the central value where the bulk of events lie, and the shadowed area the standard deviation, as measure of the variability, or how spread the observations are with respect to the mean. In this case the three observation sets, or *samples*, consist of 100 observations. In the case of χ_1, χ_2, χ_3 [...].

and the covariance is then

$$\begin{aligned}
\mathrm{Cov}(x,y) &= \frac{1}{3}\sum_{i=1}^{3}(x_i-\bar{x})(y_i-\bar{y}) \\
&= \frac{1}{3}\Big[(1-2)(2-3.67)+(2-2)(4-3.67)+(3-2)(5-3.67)\Big] \\
&= \frac{1}{3}\Big[(-1)(-1.67)+0\cdot 0.33+1\cdot 1.33\Big] \\
&= \frac{1}{3}\big[1.67+0+1.33\big] = 1\ .
\end{aligned}$$

The standard deviations are

$$\begin{aligned}
\sigma_x &= \sqrt{\frac{1}{3}\sum_{i=1}^{3}(x_i-\bar{x})^2} \\
&= \sqrt{\frac{(1-2)^2+(2-2)^2+(3-2)^2}{3}} \\
&= \sqrt{\frac{1+0+1}{3}} = \sqrt{\frac{2}{3}} \approx 0.816\ ,
\end{aligned}$$

$$\begin{aligned}
\sigma_y &= \sqrt{\frac{1}{3}\sum_{i=1}^{3}(y_i-\bar{y})^2} \\
&= \sqrt{\frac{(2-3.67)^2+(4-3.67)^2+(5-3.67)^2}{3}} \\
&= \sqrt{\frac{2.78+0.11+1.76}{3}} = \sqrt{\frac{4.65}{3}} \approx 1.24\ .
\end{aligned}$$

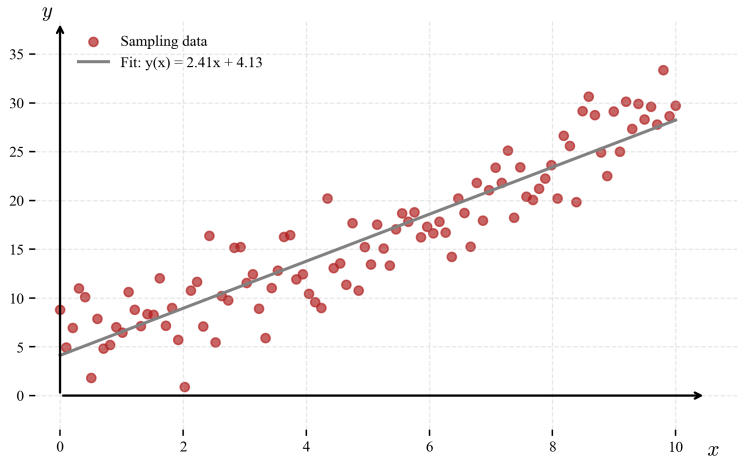


Figure 1.9: Scatter data following linear dependency. The linear function $y(x) = ax+b$ is displayed overlaying the sampling points, with parameters fitted from data.

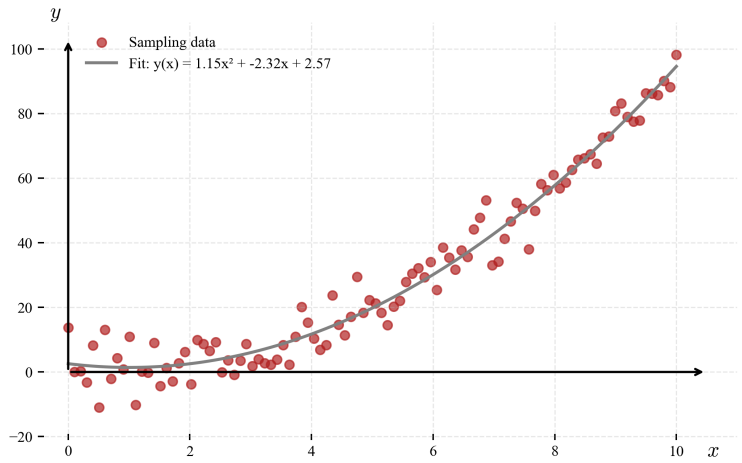


Figure 1.10: Scatter data following quadratic dependency. The quadratic function $y(x) = ax+b$ is displayed overlaying the sampling points, with parameters fitted from data.

Finally, the Pearson correlation coefficient is obtained by combining all these as we saw in (1.19)

$$\rho_{x,y} = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y} = \frac{1}{0.816 \cdot 1.24} \approx 0.99\ . \tag{1.23}$$

This positive value close to 1 indicates a strong positive linear association between x and y. We could also visualize these observations with a scatter plot, as the one displayed in Figure [...].

A note about notation: the Pearson correlation coefficient is usually denoted r for a finite sample and ρ for the idealized, complete population. Interpretation of r is quite straightforward for linear relationships: values close to ± 1 indicate a strong linear association, while values near 0 suggest weak or no linear dependency. However, Pearson's correlation has limitations: it is sensitive to outliers, captures only linear relationships, and can be misleading if the relationship is more complex.

In the context of regression, the model assumes that the variance of the error terms is constant across all values of the independent variable, an assumption known as *homoscedasticity*. In simple terms, this means that the spread of data points around the regression line should be roughly the same, regardless of the value of the independent variable. This assumption is crucial for accurate predictions and reliable statistical

inference.

$$y(x) = ax + b + \varepsilon \,, \tag{1.24}$$

where ε denotes the error term, sometimes called the *residual*. A classic demonstration of Pearson's limitations is *Anscombe's quartet*, created by the statistician Francis Anscombe in 1973 [20]. It consists of four datasets with nearly identical summary statistics—including means, variances, and correlation coefficients—yet exhibiting dramatically different distributions and patterns when plotted. This striking example highlights the necessity of visualizing data rather than relying solely on numeric summaries.

There are ways of defining correlation in more subtle cases, but they lie outside the scope of this course. Just as an example for ordinal data, where rankings rather than numeric differences matter, the *Spearman rank correlation* ρ_s is more appropriate definition. It is often used to assess monotonic relationships, was introduced by Charles Spearman in 1904 [21]. It is based on the ranks of the data rather than the raw values, making it less sensitive to outliers and non-linear relationships.

$$\rho_s = r(\mathrm{rank}(X), \mathrm{rank}(Y)). \tag{1.25}$$

This allows measurement of monotonic relationships without assuming equal spacing between values and provides a robust alternative to Pearson's r when the underlying scale is ordinal or heavily skewed..

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Chapter 2

Probability and random events

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

2.1 What is probability?

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 on games of chance, and in particular the "problem of the division of stakes," laid the groundwork for the systematic analysis of uncertain events. Years later, Jacob Bernoulli's *Ars Conjectandi* established the first classical definition of probability, providing the study of random events with mathematical clarity. Refinements by De Moivre and Laplace transformed it into a powerful analytical theory, while its true axiomatic structure only crystallised in the twentieth century with Kolmogorov's *Grundbegriffe der Wahrscheinlichkeitsrechnung* in 1933 [26].

At its heart, probability is nothing more - and nothing less - a branch of mathematics developed to describe random events, also referred to as *stochastic*. Indeed, the word "stochastic" comes from the Greek word στοχαστικός, which literally means "to guess" or "to aim." The way we describe such events, characterized by the uncertainty of their outcome, is by defining a quantity we will call $\mathbb{P}$, of probability. That quantity $\mathbb{P}$ will denote a number between 0 and 1, which reflects the degree of uncertainty, or *surprise*, with which the random event produces a specific outcome. For an event A, such as observing a heads when tossing a coin, or a given face when rolling dice, the numerical convention is written as follows,

- If I am sure A will never occur, $\mathbb{P}(A) = 0$.
- If I am sure A will always occur, $\mathbb{P}(A) = 1$.
- For anything in between, if A is *uncertain*, then $\mathbb{P}(A) \in (0, 1)$,

where the $\in$ symbol just means "belongs to". Thus, probability measures the whole span between impossibility and absolute certainty.

Consider the classic example of tossing a coin. Let H denote heads and T denote tails. If a coin is symmetric and fair, we would name the number of possible outcomes, or *sample space* $\Omega = \{H, T\}$. Those with less mathematical training may find this notation rather odd. Plain and simple, ideas such as sample space or measure space come from the underlying mathematical theory that was used to build modern probability. For the purpose of this course *sample space* will essentially mean *set of possible outcomes*.

If we are certain we will get heads heads, then $\mathbb{P}(H) = 1$ and $\mathbb{P}(T) = 0$; On the other hand, if we are certain we will get tails, the roles reverse, and then $\mathbb{P}(H) = 0$ and $\mathbb{P}(T) = 1$. In the general case, where both

outcomes can happen with equal probability, we would write

$$\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2} .$$

This assignment of 1/2 probability is not an arbitrary choice. It reflects both a symmetry of the physical system and an idealisation of experimental repetition. Indeed, the way we define probabilities for a given event A is just by computing the ratio of how many times we get that event $n(A)$, and the total number of trials N.

$$\mathbb{P}(A) = \lim_{N\to\infty} \frac{n(A)}{N} . \tag{2.1}$$

This is called the *frequentist* definition of probability, since it relies on the frequency with which each results occur. Tossing a fair coin many times, the observed frequencies of heads and tails converge towards the probabilistic assignment 1/2. The frequentist definition is built upon the idea of repetition and reproducibility. We *expect* that, if we repeat the toss many times, the number of times we get H and T will approach to a perfect half, as N increases.

$$\mathbb{P}(H) = \mathbb{P}(T) \simeq \frac{1}{2} .$$

This convergence principle is formalised in Bernoulli's Law of Large Numbers and later generalised in the Central Limit Theorem, as we will discuss in next chapter. Further approaches to the definition of probability, such as the *bayesian*, will be discussed in Chapter 5.

Beyond frequencies, probability must also obey the principle of *unitarity* or *normalisation*. This is the mathematical formalization of quite a natural intuition: at least one of the possible events must take place. By imposing that the sum of the probabilities of all mutually exclusive outcomes must equal 1, we ensure we have assigned the numerical values in a consistent way. For a finite experiment with outcomes $\{x_1, \ldots, x_n\}$, the unitarity property is written as

$$\sum_{i=1}^{n} \mathbb{P}(x_i) = 1 . \tag{2.2}$$

For a coin toss, this reduces to

$$\mathbb{P}(\mathrm{H}) + \mathbb{P}(\mathrm{T}) = \frac{1}{2} + \frac{1}{2} = 1 .$$

For a dice roll, where each face appears with a probability $P = 1/6$, we would write

$$\mathbb{P}(1) + \mathbb{P}(2) + \cdots + P(6) = \frac{1}{6} + \frac{1}{6} + \cdots + \frac{1}{6} = 1 .$$

Unitarity is one of the most fundamental properties of probability, and it will prove useful to make calculations further on this very chapter. In addition, just as a note on the formal development of all this framework, it is only once this normalisation condition imposed, that the *probability* space $(\Omega, \mathcal{F}, \mathbb{P})$ can be defined from the abstract notion of *measure* space $(\Omega, \mathcal{F}, \mu)$.

These three notions suffice for now. We have seen probability as a number that quantifies uncertainty, the frequentist definition in terms of ratio, and the idea of unitarity. Let us emphasize, though, that this formulation is actually quite recent. Even though the basic intuitions were already introduce by Bernoulli and Laplace, as we discussed, it was not until the nineteenth and twentieth centuries, that probability theory was properly formalized in the language of analysis. The idea of expectation became formally defined via the Lebesgue integral [38], stochastic processes were studied by Wiener and Doob [39], and the idea of convergence - that lied the foundations for unitarity - were explored by Borel, Cantelli, and Kolmogorov [44]. From gambling practice, probability grew into a highly abstract and powerful theory.

The way probability was mathematically defined is based on the idea of *probability space*. This may seem quite abstract at first, so let's illustrate with an example. Imagine that every experiment we perform - tossing a coin, rolling a dice, or measuring the brightness of a star - has a collection of possible outcomes. This collection is what mathematicians call the *sample space*, often written as Ω. Within this space, we

may be interested in particular groups of outcomes, such as "getting an even number" from a dice roll or "obtaining heads" from a coin toss. These groups of outcomes are what we call *events*.

Formally, a probability space is defined as a collection of three mathematical objects $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the *sample space* of possible outcomes, $\mathcal{F}$ is a collection of measurable events (for mathematicians, a σ-algebra), and $\mathbb{P}$ is a measure assigning real numbers between 0 and 1. It is simply a structured way of saying (1) the set of all possible outcomes of an experiment, (2) the events we are interested in within that set, and (3) a systematic assignment of likelihoods to those events. The more abstract terminology - σ-algebras and measures - becomes indispensable when probability theory is extended to complicated or infinite cases, but in everyday examples such as coins, dice, or cards, it suffices to remember these three key ingredients: outcomes, events, and probabilities.

The way we define and assign probabilities to random events is done in accordance with Kolmogorov's axioms, which we summarize as follows:

- **Non-negativity:** The probability of any event is never negative. Probabilities are numbers that represent likelihood, so they must satisfy $\mathbb{P}(A) \geq 0$. In simple terms, it makes no sense to say an event happens with "negative chance."
$$\mathbb{P}(A) \ \geq \ 0 \, , \qquad \forall A \in \mathcal{F} \, . \tag{2.3}$$

- **Normalisation:** The probability of the whole sample space Ω is exactly 1. This expresses the fact that "something will happen." If Ω is the complete list of possible outcomes of an experiment, then we are certain that the final outcome will be one of them.
$$\mathbb{P}(\Omega) \ = \ 1 \, . \tag{2.4}$$

- **Countable additivity:** If we have several events $A_1, A_2, \ldots$ that cannot overlap (e.g. they are mutually exclusive or disjoint), then the probability that one or another occurs is the sum of their probabilities. For instance, in rolling a die, the probability of rolling "1 or 2" is $P(1) + P(2)$ because the two outcomes cannot happen at the same time.
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \right) \ = \ \sum_{i=1}^{\infty} \mathbb{P}(A_i) \, , \qquad \text{for disjoint } A_i \in \mathcal{F} \tag{2.5}$$

where the symbol $\bigcup$ just means the *reunion* of all events A_i. Together, these three axioms form the rigorous foundation of probability theory and ensure consistency in reasoning about uncertainty.

This framework allows one to treat uncertainty with mathematical precision, applicable not only to games of chance but also to infinite-dimensional spaces, stochastic processes, and mathematical modelling of various and broad scenarios within the natural sciences. It is this generality that transformed probability theory from a tool of gamblers and actuaries into one of the central languages of modern science.

Of course, mathematics provides structure, but actual meaning requires interpretation. As we mentioned already, two main schools of interpretation emerged from this axiomatic development of probability theory. The *frequentist* definition, associated with Richard von Mises [40], identifies probability as we did in (**??**), with long-run frequency in repeated trials: it is then an objective property of the physical world, revealed through repetition. On the other hand, the *Bayesian* tradition, with origins in Thomas Bayes' posthumous essay *Towards solving a Problem in the Doctrine of Chances* back in 1763 [41], conceives probability as a measure of rational belief, updated by evidence through what we call nowadays Bayes' rule, or Bayes' theorem. Such idea was later refined by Laplace [42] and further developed by Bruno de Finetti, who further emphasized that probability expresses degrees of personal belief coherent under rational rules of betting [43].

These two traditions, frequentist and Bayesian, illuminate different facets of the same mathematical object. One interprets probability as an empirical limit of frequencies, the other as a calculus of information and belief. Both, however, are grounded in the modern axiomatic formulation: probability is a measure, and measures assign form and consistency to uncertainty.

2.2 Discrete random variables

The study of randomness begins with the idea of an *event*. An event is a possible outcome of some uncertain process: a coin landing heads, a dice showing a particular face, or a light bulb failing after a certain number of hours. To such events we attach probabilities, which are numbers between zero and one quantifying how likely they are to occur. Probability, as we defined it in the previous section, is therefore the language for describing uncertainty in a precise and quantitative way.

Yet events rarely appear in isolation. A single coin flip may be simple to describe, but when tossing many coins, rolling dice repeatedly, or counting the number of calls arriving in a given time interval, randomness begins to reveal patterns. These patterns are not entirely chaotic. Indeed, they show structure and regularities that can be modelled. To capture such patterns, we use the concept of a *random variable*. With that, he probability of obtaining a specific outcome becomes not just a single number, but a function defined over all possible values of the random variable. It is common to denote random variables with an upper case letter, such as X, and then write $\mathbb{P}(X = k)$ for the probability that X takes the value k.

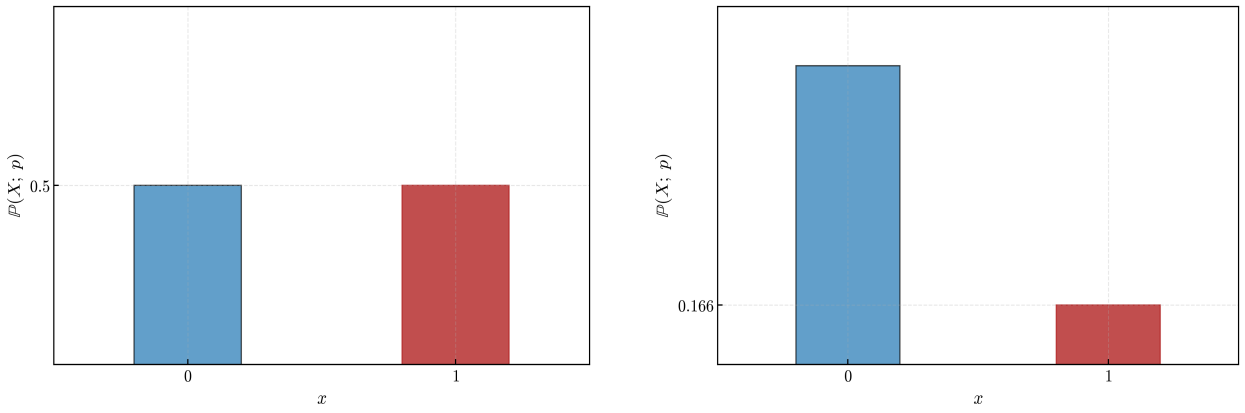
The set of all such probabilities is called a *distribution*. A distribution is hence a model, a *function*, a mathematical story about how chance unfolds in a given situation. Different phenomena give rise to different families of distributions, each with its own characteristic features. In what follows, we will introduce some of the most fundamental distributions: the Bernoulli, Binomial, Poisson, and Gaussian, among others. To begin with, we will address the so-called *discrete* random variables, and by discrete we mean they can only take a countable set of values. Tossing coins, which can be either "heads" or "tails", rolling dice, which can only yield either $1, 2, \ldots, 6$, or counting the number of calls in an hour, are some examples.

Bernoulli distribution

The Bernoulli distribution, named after Jacob Bernoulli, represents the simplest probabilistic experiment: a single trial with two possible outcomes. It is sometimes called a Bernoulli *trial*, or a Bernoulli *experiment*. A random variable X follows a Bernoulli distribution if

$$\mathbb{P}(X = 1) = p, \qquad \mathbb{P}(X = 0) = 1 - p\,. \tag{2.6}$$

Here, p is the probability of success (often denoted as $X = 1$), while $1 - p$ is the probability of failure (written as $X = 0$). The Bernoulli distribution is thus the mathematical way of describing any yes/no or true/false experiment, capturing the essence of success and failure. For a given probability of success p, a random variable X following a Bernoulli distribution is written as $X \sim \text{Bernoulli}(p)$.



(a) Coin toss.

(b) Dice roll.

Figure 2.1: Representation of the Bernoulli distribution as a histogram. The horizontal axis displays the random variable X, and the height of the bars in the vertical axis the probability for each value of X.

Binomial distribution

From the Bernoulli trial as an elementary building block arises the Binomial distribution, which describes the total number of successes in n independent Bernoulli trials. Introduced by Bernoulli and later refined by De Moivre and Laplace, the Binomial distribution became central to early probability theory. If a random variable X takes a specific value - we call it a success - with probability p, the probability of observing k successes out of n total trials, follows a Binomial distribution

$$\mathbb{P}(X = k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \, . \tag{2.7}$$

Here, n is the number of trials, p is the probability of success in each trial, and k is the total number of successes. The binomial coefficient $\binom{n}{k}$ counts the number of possible arrangements of k successes among n trials. It is normally written as $X \sim \text{Binomial}(k; n, p)$, or just $X \sim \text{Binomial}(n, p)$

Let's illustrate with an example. Imagine flipping a biased coin $n = 10$ times, where the probability of getting heads in $P = 1/2$. What is the probability of observing exactly $k = 3$ heads? By substituting in (2.7) we get

$$\mathbb{P}(X = 3;\ n = 10,\ p = 0.7) = \binom{10}{3} (0.3)^3 \ (0.7)^7 = \frac{10!}{3! \cdot 7!} \ (0.3)^3 \ (0.7)^7 \approx 0.266 \, .$$

So there is about a 26.6% chance of seeing exactly three heads.

We can compute more than just the probability of observing a single outcome. If we ask for the probability of X taking a value *less than* or *greater than* k, we would be computing a *cumulative* probability. Such quantity is computed by just adding the Binomial probabilities of all the individual cases. For example, what is the probability of observing *at most* 2 heads in 10 flips of the same biased coin? We will compute the Binomial probability of observing none, then one, then two, and add them together.

$$\mathbb{P}(X \leq 2;\ 10,\ 0.7) = \sum_{k=0}^{2} \binom{10}{k} (0.3)^k \ (0.7)^{10-k} \, .$$

Compute each term:

$$\begin{aligned}
\mathbb{P}(X = 0;\ 10,\ 0.7) &= \binom{10}{0} (0.3)^0 \ (0.7)^1 0 = 1 \cdot (0.7)^{10} \approx 0.028, \\
\mathbb{P}(X = 1;\ 10,\ 0.7) &= \binom{10}{1} (0.3)^1 \ (0.7)^9 = 10 \cdot 0.3 \cdot (0.7)^9 \approx 0.121, \\
\mathbb{P}(X = 2;\ 10,\ 0.7) &= \binom{10}{2} (0.3)^2 \ (0.7)^8 = 45 \cdot 0.09 \cdot (0.7)^8 \approx 0.233.
\end{aligned}$$

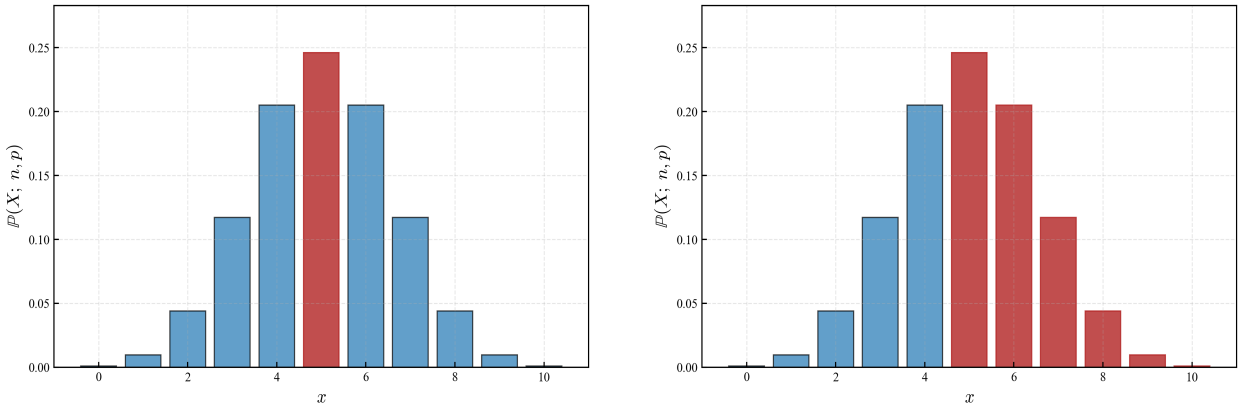
Thus $\mathbb{P}(X \leq 2) \approx 0.028 + 0.121 + 0.233 = 0.382$, so there is about a 38% chance of observing two or fewer heads.

Poisson distribution

The Poisson distribution, formulated by Siméon-Denis Poisson in *Recherches sur la probabilité des jugements* back in 1837 [33], models the occurrence of rare events over a fixed interval. If we ask what is the probability of observing a specific number k of events, given a certain average λ, that probability follows a Poisson distribution

$$\mathbb{P}(X = k; \lambda) = \frac{e^{-\lambda} \ \lambda^k}{k!} \, . \tag{2.8}$$

Here, λ is the expected number of events in the interval, while k is the observed count of events. It is written $X \sim \text{Poisson}(k;\ \lambda)$, or just $X \sim \text{Poisson}(\lambda)$. The Poisson distribution models a very broad set of phenomena, ranging from telephone call arrivals to radioactive decay.



(a) Probability of single event $\mathbb{P}(X = k)$.

(b) Cumulative probability $\mathbb{P}(X > k)$.

Figure 2.2: Representation of the Binomial distribution as a histogram. The horizontal axis displays the random variable X, and the height of the bars in the vertical axis the probability for each value of X.

Let's illustrate with an example. If the average of meteors falling per night in a given area is $\lambda = 4$, what would be the probability of observing exactly 6 meteors in one night? By substituting in (2.8) we get

$$\mathbb{P}(X = 6;\ \lambda = 4) = \frac{e^{-4} \cdot 4^6}{6!} = \frac{0.0183 \cdot 4096}{720} \approx 0.104\ .$$

So there is about a 10.4% chance of exactly six meteors.

Again, we could ask a cumulative probability. What is the probability of observing *less than* 2 meteors in that same place?

$$\mathbb{P}(X \leq 2;\ \lambda = 4) = \sum_{k=0}^{2} \frac{e^{-4} \cdot 4^k}{k!}\ . \tag{2.9}$$

Compute each term:

$$\begin{aligned}
\mathbb{P}(X = 0;\ 4) &= e^{-4}\ \frac{4^0}{0!} = e^{-4} \cdot 1 \approx 0.018\ , \\
\mathbb{P}(X = 1;\ 4) &= e^{-4}\ \frac{4^1}{1!} = e^{-4} \cdot 4 \approx 0.073\ , \\
\mathbb{P}(X = 2;\ 4) &= e^{-4}\ \frac{4^2}{2!} = e^{-4} \cdot 8 \approx 0.146\ .
\end{aligned}$$

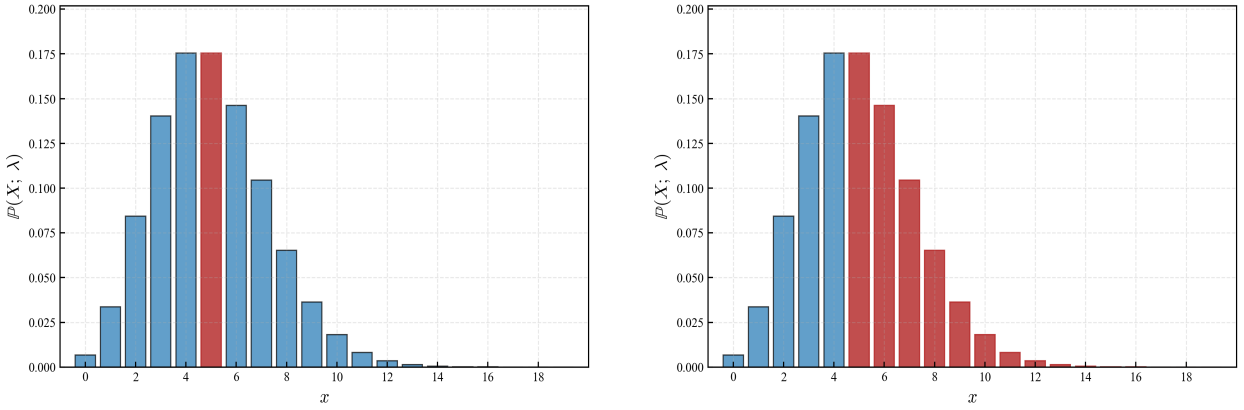
Thus $\mathbb{P}(X \leq 2) \approx 0.018 + 0.073 + 0.146 = 0.237$, so there is about a 24% chance of seeing two or fewer meteors.

Discrete uniform distribution

Finally, the discrete uniform distribution is in a way the simplest one. It reflects the symmetry inherent in classical games of chance, assigning equal likelihood to each outcome. It was historically motivated by fair dice and lotteries, and it is attributed to Cardano's *Liber de Ludo Aleae*, back in the late sixteenth century [34]. For X uniformly distributed over $\{1, 2, \ldots, n\}$ possible outcomes,

$$\mathbb{P}(X = k;\ n) = \frac{1}{n}. \tag{2.10}$$

Here, n is the number of equally possible outcomes, and each k from 1 to n is assigned the same probability. We see that the $1/2$ probabilities for heads and tails in a fair coin, or the flat $1/6$ for the faces of the dice, are just some example of the discrete uniform distribution.



(a) Probability of single event $\mathbb{P}(X = k)$.

(b) Cumulative probability $\mathbb{P}(X = k)$.

Figure 2.3: Representation of the Poisson distribution as a histogram. The horizontal axis displays the random variable X, and the height of the bars in the vertical axis the probability for each value of X.

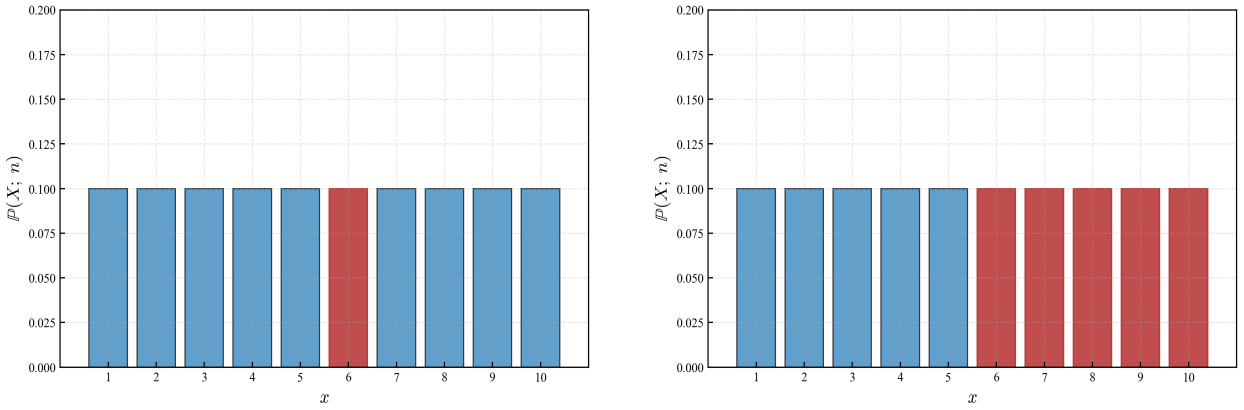
The examples here are quite trivial. Obtaining $X = 3$ on a fair six-sided dice just yields

$$\mathbb{P}(X = 3;\ n = 6) = \frac{1}{6}\,.$$

Let's go for the cumulative case. Computing the probability of rolling a number *less than or equal* to 4, since all outcomes are equally likely,

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

This captures a natural intuition: when all outcomes are equally probable, cumulative probabilities simply count favourable cases over total cases.



(a) Probability of single event $\mathbb{P}(X = k)$.

(b) Cumulative probability $\mathbb{P}(X > k)$.

Figure 2.4: Representation of the discrete Uniform distribution as a histogram. The horizontal axis displays the random variable X, and the height of the bars in the vertical axis the probability for each value of X.

2.3 Continuous random variables

So far we have just discussed the discrete case, where random variables can only take a countable set of values: tossing coins, rolling dice, counting the number of calls in an hour, or the number of defective

items in a sample. Yet many of the most important phenomena in science and daily life are not discrete, but *continuous*. A person's height is not limited to a few centimeters, the waiting time at a bus stop is not restricted to whole minutes, the voltage in a circuit does not jump in steps, etc. Instead, these quantities can vary *smoothly*, taking infinitely many possible values, in what we call a continuous interval.

To model such phenomena, we need a new kind of probability distribution. Imagine we try to compute the probability of measuring a specific temperature in a room. For any value, if we try to apply the frequentist definition of probability, the probability of obtaining exact value (say, "exactly 25 degrees") is exactly zero, since there are infinitely many possible values. Instead, we will define a new mathematical object, a *probability density function* (pdf), denoted $f(x)$. While $f(x)$ itself is not itself a probability, its integral over an interval gives the probability of the random variable falling in that interval:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx\ . \tag{2.11}$$

the assign probabilities to *intervals*, such as "between 20 and 25 degrees". This shift from individual points to continuous densities is profound: it allows us to describe uncertainty with infinite resolution. Continuous distributions thus play a central role in physics, engineering, biology, economics, and the social sciences. They provide not only mathematical elegance, but also a remarkably accurate reflection of how the world behaves when measured with fine instruments.

Continuous uniform distribution

The continuous uniform distribution generalizes the principle of indifference, elegantly articulated by Laplace [32]. By assigning equal probability density across an interval, it models complete ignorance about a continuous quantity. For $X \sim \text{Uniform}(a, b)$,

$$f(x) = \frac{1}{b-a} \quad (a \leq x \leq b)\ . \tag{2.12}$$

The parameters a and b mark the left and right endpoints of the interval of possible values. Every subinterval of the same length is equally likely, because the density is constant at $1/(b-a)$. The mean is the midpoint $(a+b)/2$, reflecting the symmetry; the variance depends only on the width $b-a$, becoming larger as the interval widens.

Example 1 (basic properties). A bus is equally likely to arrive at any time between 5 and 15 minutes past the hour. Then $X \sim \text{Uniform}(5, 15)$, so

$$f(x) \;=\; \frac{1}{15-5} \;=\; 0.1 \quad \text{for } 5 \leq x \leq 15. \tag{2.13}$$

The constant density expresses that no minute in the window $[5, 15]$ is preferred; the average waiting time is exactly the midpoint.

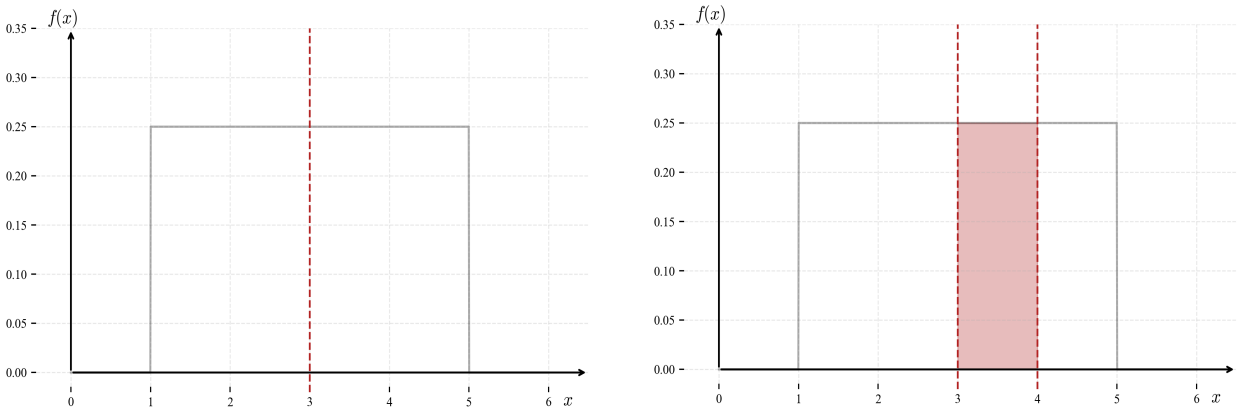
Example 2 (cumulative). What is the probability the bus arrives between 5 and 13 minutes past the hour?

$$\mathbb{P}(5 \leq X \leq 13) \;=\; \int_5^{13} 0.1\,dx. \tag{2.14}$$

Compute the integral:

$$\int_5^{13} 0.1\,dx \;=\; 0.1 \times (13-5) \;=\; 0.1 \times 8 \;=\; 0.8. \tag{2.15}$$

With a uniform density, probabilities are proportional to lengths. The interval $[5, 13]$ covers four-fifths of the full window $[5, 15]$, so the probability is 0.8.



(a) Probability density $f(x)$.

(b) Cumulative probability $\int_a^b f(x)\,dx$.

Figure 2.5: Representation of the Uniform distribution as a continuous function. The horizontal axis displays the random variable x.

Exponential distribution

The exponential distribution arises naturally in contexts where events occur randomly over time, such as radioactive decay, system failures, or arrivals in a queue. Its distinctive property is sometimes referred as *memoryless*: the probability of an event occurring in the next interval does not depend on how much time has already passed. For $X \sim \text{Exp}(\lambda)$,

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0\,. \tag{2.16}$$

The parameter λ is the average *rate* of occurrence (events per unit time). Large λ means events happen frequently and waiting times are typically short; small λ means events are rare and waits are longer. The hallmark "memoryless" property means: given that no event has occurred up to time t, the remaining waiting time is distributed exactly as if we were starting afresh.

Example 1 (probability within a time window). A system has failure rate $\lambda = 0.2$ per hour. What is the probability it fails within 3 hours?

$$\mathbb{P}(X \leq 3) \;=\; \int_0^3 0.2\, e^{-0.2x}\, dx. \tag{2.17}$$

Compute integral

$$\int 0.2\, e^{-0.2x}\, dx \;=\; -\,e^{-0.2x} + C. \tag{2.18}$$

Evaluate from 0 to 3:

$$\mathbb{P}(X \leq 3) \;=\; \big[-\,e^{-0.2x}\big]_0^3 \;=\; \big(-\,e^{-0.6}\big) - \big(-\,e^{0}\big) \;=\; 1 - e^{-0.6}. \tag{2.19}$$

Numerically,

$$e^{-0.6} \approx 0.5488 \quad \Rightarrow \quad \mathbb{P}(X \leq 3) \approx 1 - 0.5488 \approx 0.4512. \tag{2.20}$$

There is about a 45% chance of failure within 3 hours; the complement ($\approx 55\%$) is the chance it lasts longer than 3 hours.

Example 2 (cumulative over an interval). Calls arrive with rate $\lambda = 0.5$ per minute. What is the probability the waiting time until the next call is between 2 and 5 minutes?

$$\mathbb{P}(2 \leq X \leq 5) \;=\; \int_2^5 0.5\, e^{-0.5x}\, dx. \tag{2.21}$$

Compute integegral:

$$\int 0.5\, e^{-0.5x}\, dx \;=\; -\,e^{-0.5x} + C. \tag{2.22}$$

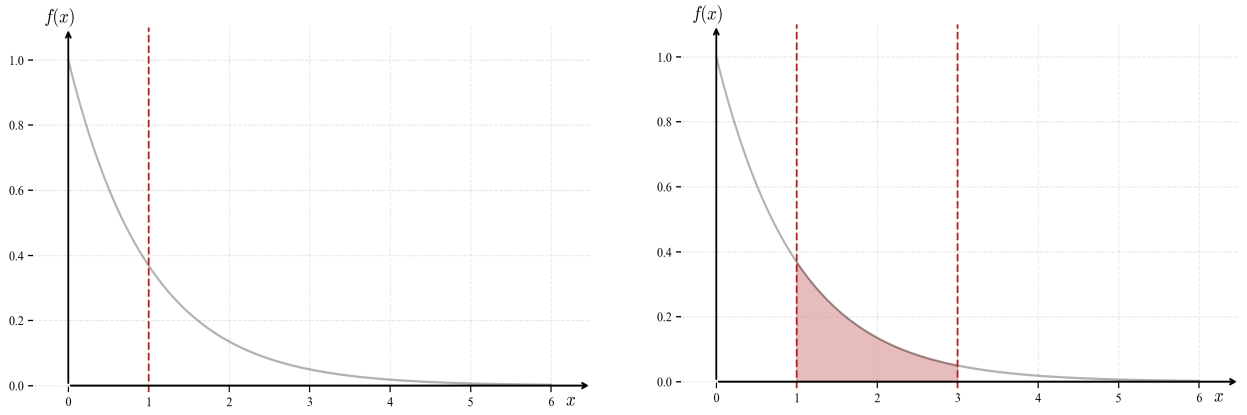
Evaluate:

$$\mathbb{P}(2 \le X \le 5) \;=\; \big[-\,e^{-0.5x}\big]_2^5 \;=\; e^{-1} \;-\; e^{-2.5}. \tag{2.23}$$

Approximate numerically:

$$e^{-1} \approx 0.3679, \qquad e^{-2.5} \approx 0.0821, \qquad \Rightarrow \quad \mathbb{P}(2 \le X \le 5) \approx 0.3679 - 0.0821 \approx 0.2858. \tag{2.24}$$

There is roughly a 29% chance the next call arrives between 2 and 5 minutes. Exponential tails decrease smoothly, so longer waits are progressively less likely.



(a) Probability density $f(x)$.

(b) Cumulative probability $\int_a^b f(x)\, dx$.

Figure 2.6: Representation of the Exponential distribution as a continuous function. The horizontal axis displays the random variable x.

Gaussian distribution

The Gaussian (normal) distribution, introduced rigorously by Carl Friedrich Gauss in his 1809 study of least squares [35], formalizes the archetype of natural variation. Prefigured by De Moivre, it underlies the Central Limit Theorem and pervades modern statistics:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\!\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{2.25}$$

The parameter μ is the centre (mean) of the bell curve and σ measures spread: most values lie within a few σ of μ, with probabilities tapering smoothly in the tails. The distribution is symmetric about μ, so values equally far above and below the mean are equally likely.

Example 1 (upper tail). Suppose adult heights are approximately normal with $\mu = 170$ cm and $\sigma = 10$ cm. What is the probability that a randomly chosen person is taller than 180 cm?

$$\mathbb{P}(X > 180) \;=\; \mathbb{P}\!\left(\frac{X-\mu}{\sigma} > \frac{180-170}{10}\right) \;=\; \mathbb{P}(Z > 1), \tag{2.26}$$

where $Z \sim \mathcal{N}(0,1)$ is standard normal. Using the standard normal cdf Φ,

$$\mathbb{P}(Z > 1) \;=\; 1 - \Phi(1). \tag{2.27}$$

From tables (or software),

$$\Phi(1) \approx 0.8413 \quad \Rightarrow \quad \mathbb{P}(X > 180) \approx 1 - 0.8413 \approx 0.1587. \tag{2.28}$$

About 16% of individuals exceed 180 cm in this model—consistent with the bell shape: one standard deviation above the mean lies in the upper tail.

Example 2 (between two bounds). What is the probability that a height lies between 160 cm and 185 cm?

$$\mathbb{P}(160 \le X \le 185) \;=\; \mathbb{P}\left(\frac{160-170}{10} \le \frac{X-\mu}{\sigma} \le \frac{185-170}{10}\right) \;=\; \mathbb{P}(-1 \le Z \le 1.5). \tag{2.29}$$

Express this via Φ:

$$\mathbb{P}(-1 \le Z \le 1.5) \;=\; \Phi(1.5) - \Phi(-1). \tag{2.30}$$

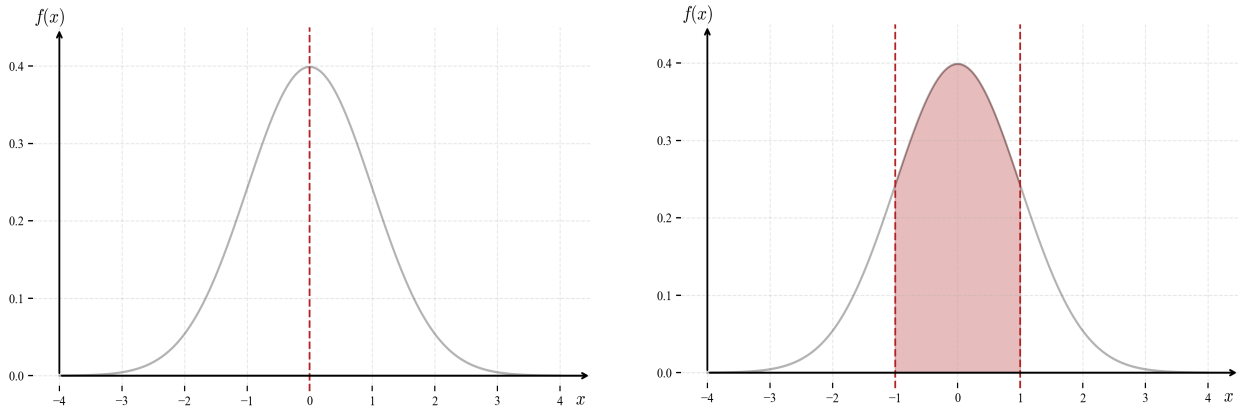
Use symmetry $\Phi(-z) = 1 - \Phi(z)$ (or a table value for -1):

$$\Phi(1.5) \approx 0.9332, \qquad \Phi(-1) \approx 0.1587, \tag{2.31}$$

so

$$\mathbb{P}(160 \le X \le 185) \;\approx\; 0.9332 - 0.1587 \;=\; 0.7745. \tag{2.32}$$

Roughly 77% of heights fall in this range. The interval spans slightly more than one standard deviation below the mean up to one and a half above, capturing most of the bell curve's mass.



(a) Probability density $f(x)$.

(b) Cumulative probability $\int_a^b f(x)\,dx$.

Figure 2.7: Representation of the Gaussian distribution as a continuous function. The horizontal axis displays the random variable x.

2.4 Expectation values

An essential concept in probability theory and statistics is the *expected value*, sometimes called the *mathematical expectation*. Intuitively, the expected value represents the long-run average outcome of a random experiment if it were repeated many times. This provides a systematic way to capture the "center" of a distribution of outcomes.

The expected value is best understood as a weighted average, where each possible outcome of a random variable is multiplied by the probability of its occurrence. For a discrete random variable, we sum over all possible values; for a continuous random variable, we integrate against the probability density. Beyond the mean, we can also measure the *spread* of the distribution: the variance, defined as the expected squared deviation from the mean. Higher moments extend this principle, capturing skewness and kurtosis.

Formally, if X is a random variable, its expectation is defined as follows.

$$\mathbb{E}[X] = \sum_x x\, p(x), \qquad \text{if } X \text{ is discrete,}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,dx, \qquad \text{if } X \text{ is continuous,}$$

where $p(x)$ is a probability mass function and $f(x)$ a probability density function. The variance is given by

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

Warmup Examples

Discrete case. Suppose X represents the outcome of rolling a fair die. Then

$$\mathbb{E}[X] = \sum_{k=1}^{6} k \cdot \tfrac{1}{6} = \frac{1+2+3+4+5+6}{6} = 3.5.$$

The variance is

$$\operatorname{Var}(X) = \frac{1}{6}\sum_{k=1}^{6}(k-3.5)^2 = \frac{35}{12}.$$

Continuous case. Suppose $X \sim \text{Uniform}(0,1)$. Then

$$\mathbb{E}[X] = \int_0^1 x\,dx = \tfrac{1}{2}, \qquad \operatorname{Var}(X) = \int_0^1 (x-\tfrac{1}{2})^2\,dx = \tfrac{1}{12}.$$

Distributions

Bernoulli. If $X \sim \text{Bernoulli}(p)$,

$$\mathbb{P}(X=1) = p, \quad \mathbb{P}(X=0) = 1-p, \quad \mathbb{E}[X] = p, \quad \operatorname{Var}(X) = p(1-p).$$

Binomial. If $X \sim \text{Binomial}(n,p)$,

$$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \mathbb{E}[X] = np, \quad \operatorname{Var}(X) = np(1-p).$$

Poisson. If $X \sim \text{Poisson}(\lambda)$,

$$\mathbb{P}(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad \mathbb{E}[X] = \lambda, \quad \operatorname{Var}(X) = \lambda.$$

Discrete Uniform. If $X \sim \text{Uniform}\{1,2,\ldots,n\}$,

$$\mathbb{P}(X=k) = \frac{1}{n}, \qquad \mathbb{E}[X] = \frac{n+1}{2}, \qquad \operatorname{Var}(X) = \frac{n^2-1}{12}.$$

Continuous Uniform. If $X \sim \text{Uniform}(a,b)$,

$$f(x) = \frac{1}{b-a} \quad (a \le x \le b), \qquad \mathbb{E}[X] = \frac{a+b}{2}, \qquad \operatorname{Var}(X) = \frac{(b-a)^2}{12}.$$

Exponential. If $X \sim \text{Exp}(\lambda)$,

$$f(x) = \lambda e^{-\lambda x}, \quad x \ge 0, \qquad \mathbb{E}[X] = \frac{1}{\lambda}, \qquad \operatorname{Var}(X) = \frac{1}{\lambda^2}.$$

Normal. If $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad \mathbb{E}[X] = \mu, \qquad \mathrm{Var}(X) = \sigma^2.$$

More generally, the expectation provides the foundation for defining the *moments* of a random variable. For a random variable X, the n-th moment about the mean is

$$\mu_n = \mathbb{E}[(X - \mathbb{E}[X])^n].$$

- The **first moment** $\mu_1 = 0$ by construction, but the raw moment $\mathbb{E}[X]$ is the *mean*.
- The **second central moment** μ_2 is the *variance*, measuring spread.
- The **third central moment** μ_3 leads to the *skewness*, a normalized measure of asymmetry:

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}.$$

- The **fourth central moment** μ_4 leads to the *kurtosis*, a normalized measure of tail heaviness:

$$\gamma_2 = \frac{\mu_4}{\mu_2^2}.$$

Thus the mean, variance, skewness, and kurtosis can all be seen as successive refinements in describing a distribution, moving from its center to its spread, asymmetry, and tail behaviour.

Moments were systematically employed by Pafnuty Chebyshev in the mid-nineteenth century to prove inequalities and the weak law of large numbers. They later became central to the work of Karl Pearson, who introduced skewness and kurtosis as moment-based measures of distributional shape [37]. The use of the moment-generating function, $M_X(t) = \mathbb{E}[e^{tX}]$, was an innovation of the early twentieth century, providing a bridge between analysis and probability. In summary, moments capture essential geometric features of probability distributions, serving as both summary statistics and analytical tools. Their historical trajectory, from Huygens through Chebyshev to Pearson, mirrors the development of probability theory from gambling practice to rigorous science.

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Chapter 3

Parameter estimation

Numbers have an important story to tell.
They rely on us to give them a voice.
— Florence Nightingale

3.1 Prediction vs inference

In the previous chapters we have introduced the mathematical theory of probability. We have developed a series of tools, a *theory*, which enables us to make predictions in stochastic processes. But, contrary to what is normally explain in introductory courses, science is not always headed in the theory - prediction - experiment direction. There can be cases, as we will soon see, where hypothesis are formulated for a given phenomena, and no prediction is made. In such cases, it is from measurement that we will try to see, or *infer* if a given set of assumptions are compatible with the obtained data. Indeed, most modern data analysis and hypothesis testing lie in the *inferential* statistics, rather than *predictive* probability [...].

Inference seeks to explain why and how variables relate. The key idea is causality and interpretability: given a some set of observations, inference aims to answer questions such as: Does smoking cause lung cancer, or is the correlation due to other confounding factors? How does an increase in temperature affect ice cream sales? What are the most significant predictors of house prices? The difference between prediction and inference has been a topic of interest in statistics and data science for centuries. While both concepts involve drawing conclusions from data, their goals, methodologies, and historical development differ significantly [...].

The roots of inference trace back to classical statistics, particularly the work of Laplace (1749–1827) Gauss (1777–1855), who developed probability theory and the method of least squares. Their work laid the foundation for statistical inference, which aims to understand relationships between variables and make generalizable conclusions about populations from samples. For example, Laplace used probability theory to estimate the population of France, introducing Bayesian inference, which provides a framework for updating beliefs based on observed data. Gauss contributed the normal distribution and least squares estimation, which became essential for making inferences about unknown parameters.

Statistical techniques such as hypothesis testing, confidence intervals, and regression analysis aim to understand and describe these relationships. The emphasis lies on estimating parameters and determining statistical significance rather than simply making accurate predictions. A classic example is Ronald Fisher (1890–1962), who developed maximum likelihood estimation (MLE) to infer parameters of probability distributions. [...].

Prediction focuses on accuracy and generalization rather than explaining causality. The goal is to create a model that performs well on new, unseen data, even if the underlying relationships between variables

are not fully understood. For example, in modern deep learning, neural networks can recognize faces with high accuracy but offer little interpretability in how they make decisions. Unlike inference, which aims to understand why a pattern exists, prediction is about making the best possible guess given the available data. Focus shifted from understanding relationships to optimizing models that generalize well to unseen data. In 2001, Leo Breiman, in his seminal paper "Statistical Modelling: The Two Cultures," highlighted the distinction, arguing that traditional statistics emphasized inference, whereas modern machine learning prioritized prediction.

3.2 Parameters, variables, statistics

A *parameter* is a number that describes the population, theoretical quantity, cannot be observed. A *statistic* is a number calculated from data, used to describe / summarie data [...]. Another key difference we will discuss now, and quite a subtle one from the mathematical perspective, is that one between a *variable* and a *parameter*. Consider the example of a binomial experiment, e.g. tossing coins and asking for the probability of measuring a specific number of heads. There, we would write it as

$$P(x; n, p) = \binom{n}{k} p^x (1-p)^{n-x}, \tag{3.1}$$

where n is the number of trials and p is the probability of success.

In our previous examples, we have treated just x as our variable of interest, but we could think about P as a function of three independent variables. The number of times we want to observe heads, the total number of trials, and the probability of success for each toss. Normally, we will call *parameters*, to all these variables we will freeze for the purpose of our calculations, and either consider them either known, or fit them from data [...].

3.3 The Law of Large Numbers

The *Law of Large Numbers* (LLN) is a fundamental theorem in probability theory that formalizes the idea that averages computed from large samples tend to stabilize, and approximate the expected value of the underlying random process. In essence, it states that as the number of trials increases, the sample mean of a sequence of random observations converges to the population mean. This result justifies the very concept of statistical average, and forms the basis for almost all of empirical science, data analysis, and applied statistics.

The Law of Large Numbers expresses a powerful kind of regularity that emerges from randomness. Suppose you flip a fair coin many times. While the outcome of any single flip is uncertain, the proportion of heads observed in the long run should converge to 0.5. The underlying reason for this stabilization lies in the independence of trials and the additivity of probability: large samples "dilute" random fluctuations and reveal the structure of the distribution. The LLN can be understood as a consequence of the averaging process: the sum of independent random deviations around the mean tends to cancel out as the sample size increases, allowing the average to settle around its expected value.

The earliest version of the Law of Large Numbers was proved by Jacob Bernoulli in 1713, posthumously published in his *Ars Conjectandi*. He proved that the proportion of successes in a sequence of independent Bernoulli trials converges in probability to the success probability. This became known as the *Weak Law of Large Numbers* (WLLN). In the 19th and 20th centuries, the LLN was generalized and refined. Russian mathematician Pafnuty Chebyshev extended the law beyond Bernoulli trials by introducing moment inequalities. Later, it was again Kolmogorov who provided a stronger formulation, the *Strong Law of Large Numbers* (SLLN) - which guarantees almost sure convergence of the sample mean to the expected value under broader conditions.

The Law of Large Numbers states that as the sample size increases, the sample mean approaches the expected value. Formally, if $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables with expected value $\mathbb{E}[X] = \mu$, then:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{as } n \to \infty. \tag{3.2}$$

Consider flipping a fair coin multiple times. The proportion of heads observed converges to 0.5 as the number of flips increases. This illustrates that the observed average stabilizes around the theoretical probability.

The Weak Law can be phrased as follows. Let $x_1, x_2, \ldots, x_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with finite expected value $\mu = \mathbb{E}[X_i]$. We could define the sample mean as we just learned in previous section:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} n_i. \tag{3.3}$$

Then, the *Weak Law of Large Numbers* states that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) = 0. \tag{3.4}$$

This means that sample mean $\overline{x}_n$ converges to the true value μ ias we increase the number of trials, $n \to \infty$.

The Strong Law: under the same conditions, and with the additional requirement that $mathbbE[|x_i|] < \infty$, the **Strong Law of Large Numbers** states that:

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1. \tag{3.5}$$

This is a stronger mode of convergence known as *almost sure convergence*, meaning that the sequence $\overline{X}_n$ converges to μ with probability 1. Kolmogorov proved this version in 1933 using tools from measure theory, laying the foundation for modern probability theory.

Example: tossing coins
Let X_i be a fair coin flip, coded as $X_i = 1$ for heads and $X_i = 0$ for tails. Then:

$$\mu = \mathbb{E}[X_i] = 0.5.$$

Now flip a coin $n = 10$, then $n = 50$, then $n = 500$ times. At each step, compute the sample mean:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

You will observe that $\overline{X}_n$ fluctuates at first, but as n increases, the value stabilizes near 0.5. This is a manifestation of the Law of Large Numbers: although the outcome of individual trials is unpredictable, the average of many trials is highly predictable and converges toward the expected value.

Example: rolling dice. Suppose we roll a fair six-sided die multiple times. The expected value of a roll is:

$$\mu = \mathbb{E}[X] = \frac{1+2+3+4+5+6}{6} = 3.5.$$

As we roll more dice, the sample mean of observed values gets closer to 3.5.

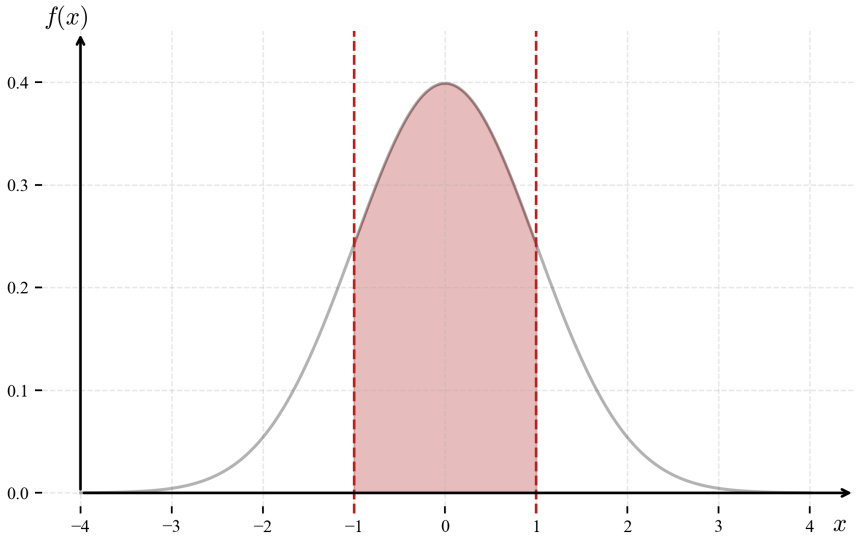


Figure 3.1: Representation of the law or large numbers. The sample mean tends to the population mean as the number of rolls n increases.

3.4 The Central Limit Theorem

The *Central Limit Theorem* (CLT) is one of the most fundamental results of probability theory. It asserts that, under general conditions, the sum (or average) of a large number of independent random variables behaves approximately like a normally distributed variable, even if the original variables themselves are not normally distributed. This surprising and powerful result provides the mathematical explanation for the pervasive appearance of the bell-shaped curve in natural and social phenomena, and it serves as the theoretical backbone of most inferential statistics.

The CLT emerges from the cumulative effect of many small, independent random influences. Imagine measuring something like the height of individuals, the result of which is shaped by many genetic, nutritional, and environmental factors. These influences may individually follow different distributions - some skewed, some bounded - but together, their additive or average effect tends to "smooth out" into a normal distribution. This smoothing occurs because the distribution of the sum is shaped by the convolution of the individual distributions, and repeated convolutions (under certain conditions) tend to produce a Gaussian shape—a phenomenon akin to a probabilistic version of the law of large numbers.

The intuitive reason for this is rooted in the *independence* of the variables and the *averaging effect* of the sum. Each random variable contributes a little "noise" to the final result, and the accumulation of many such noises results in a distribution that reflects the most probable outcome: symmetrically clustered around the mean, with tails governed by how much variability the individual sources introduce.

The first instance of a central limit-like result appears in the work of Abraham de Moivre in 1733, who demonstrated that the binomial distribution could be approximated by the normal distribution under certain conditions. Then Laplace extended this in 1810 to more general discrete distributions. These early forms assumed identically distributed, bounded, and often symmetric variables.

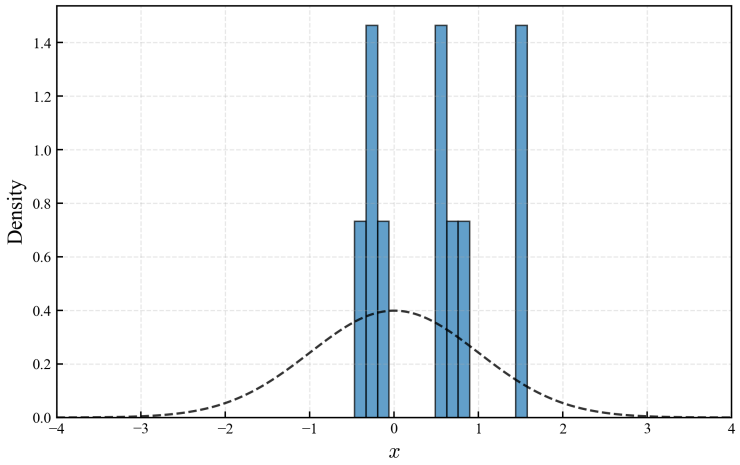


Figure 3.2: Histogram representing a set of observations and the central limit theorem. The sample mean follows a gaussian distribution as the sample size n increases.

A more general and rigorous version came from Russian mathematician Aleksandr Lyapunov in 1901, who removed many of these restrictions and provided a sufficient condition for convergence using higher-order moments. Later, J. W. Lindeberg refined the result further introducing the now-famous *Lindeberg condition*, which allows for non-identically distributed variables and still guarantees convergence under the right circumstances.

Formal Statement (Classical Version): Let $x_1, x_2, \ldots, x_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with some finite expected value $\mu = \mathbb{E}[X_i]$ and finite, positive variance $\sigma^2 = \mathrm{Var}(X_i)$. We can define the sample mean as we did in previous section:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.6}$$

Then the Central Limit Theorem (in its classical form, as due to Lyapunov, 1901) states that the distribution of the *normalized* - or *standardized* - sample mean

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \tag{3.7}$$

converges to the standard normal distribution $\mathcal{N}(0,1)$ as $n \to \infty$. That is,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq z\right) = \Phi(z), \quad \text{for all } z \in \mathbb{R}, \tag{3.8}$$

where $\Phi(z)$ is the cumulative distribution function of the standard normal distribution:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt. \tag{3.9}$$

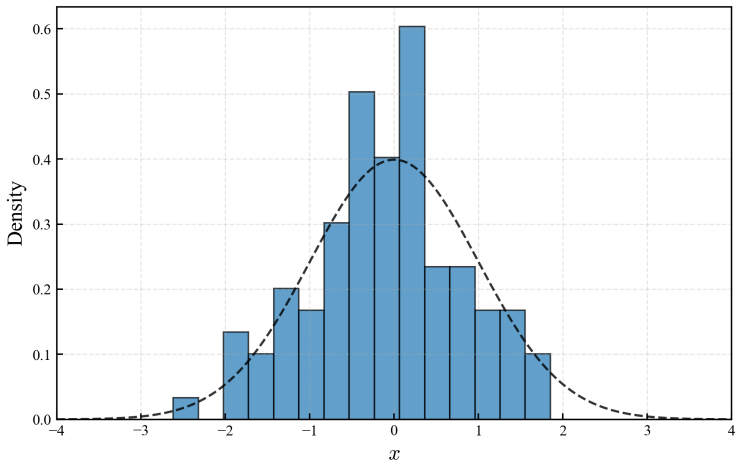


Figure 3.3: Histogram representing a set of observations and the central limit theorem. The sample mean follows a gaussian distribution as the sample size n increases.

This form of convergence—called *convergence in distribution* - means that the shape of the probability distribution of Z_n approaches that of a standard normal variable as the sample size increases, regardless of the original distribution of the X_i, provided they are i.i.d. with finite mean and variance.

Example
To illustrate the CLT in action, consider the following basic pen-and-paper experiment:
Let each X_i be a discrete uniform random variable on $\{1,2,3,4,5,6\}$, modeling a fair die roll. Then:
Compute expected, true mean value

$$\mu = \mathbb{E}[X_i] = \frac{1+2+3+4+5+6}{6} = 3.5,$$

Compute variance operator

$$\sigma^2 = \mathrm{Var}(X_i) = \frac{1}{6}\sum_{k=1}^{6}(k-3.5)^2 = \frac{35}{12} \approx 2.9167.$$

Now simulate $n = 30$ rolls, compute the sample mean $\overline{X}_{30}$, and normalize:

$$Z_{30} = \frac{\sqrt{30}(\overline{X}_{30} - 3.5)}{\sqrt{35/12}}.$$

Repeat this experiment several times (either with real dice or a random number generator). You will find that the histogram of the resulting Z_{30} values increasingly resembles the standard normal distribution. This is the Central Limit Theorem in action: even though each die roll is far from normally distributed (it's discrete and uniform), the average of many such rolls behaves approximately normally.

Conclusion, the Central Limit Theorem provides a bridge between the randomness of individual observations and the predictability of averages. It explains why normality arises in so many empirical settings and justifies the widespread use of the normal distribution in statistical methods. As sample sizes grow, the shape of the distribution of sample means stabilizes around the normal curve—allowing us to invoke Gaussian models even in non-Gaussian worlds.

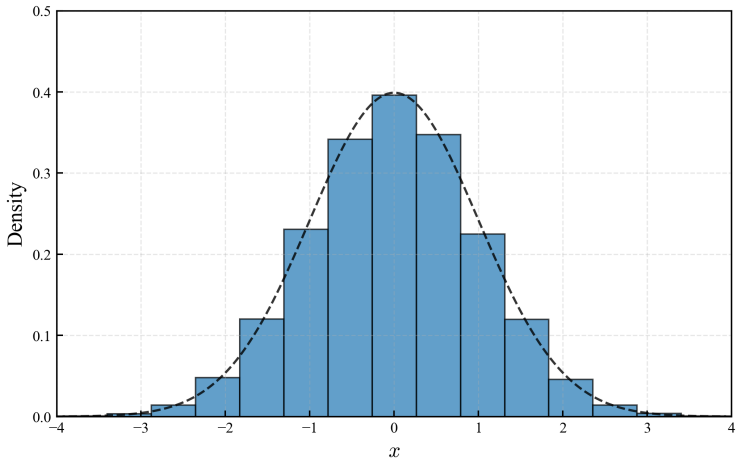


Figure 3.4: Histogram representing a set of observations and the central limit theorem. The sample mean follows a gaussian distribution as the sample size n increases.

3.5 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is one of the most fundamental methods in statistical inference. It provides a systematic way to estimate unknown parameters of a statistical model by selecting the parameter values that make the observed data most probable. Conceptually intuitive and mathematically powerful, MLE serves as a unifying framework across many areas of statistics, from simple models to complex machine learning systems.

Motivation and Intuition

Suppose we observe some data that we believe arises from a probabilistic model depending on unknown parameters. The central idea of MLE is to reverse the process: rather than asking "what data might result from given parameters?", we ask "which parameters are most likely to have generated the data we actually observed?"

More formally, we treat the likelihood—the probability (or probability density) of the observed data given a parameter—as a function of the parameter. We then choose the value of the parameter that maximizes this function. This value is called the *maximum likelihood estimate*.

This approach aligns with our everyday reasoning: if you observe a coin come up heads 90 out of 100 times, the parameter value (e.g., probability of heads) that best explains this data is likely around 0.9, not 0.5.

Historical Background

The method of maximum likelihood was introduced by **Ronald A. Fisher** in a series of seminal papers beginning in 1912 and formally developed in 1922. Fisher championed likelihood as a key concept in statistical inference, distinguishing it from Bayesian methods and providing it with a firm theoretical foundation. He also showed that under regularity conditions, MLEs are consistent, asymptotically normal, and efficient—properties that make them highly desirable estimators.

Formal Definition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with probability density function (pdf) or probability mass function (pmf) $f(x; \theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$ is the (possibly multivariate) parameter to be estimated.

The **likelihood function** is defined as:

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta),$$

viewed as a function of θ, with the observed data $x_1, \ldots, x_n$ held fixed.

The **maximum likelihood estimator** (MLE) $\hat{\theta}_{\text{MLE}}$ is the value of θ that maximizes this likelihood function:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n).$$

Often it is more convenient to maximize the **log-likelihood function**, since the logarithm is a monotonic transformation and turns the product into a sum:

$$\ell(\theta) = \log L(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i; \theta).$$

Then the MLE is equivalently given by:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \ell(\theta).$$

A Simple Example: Estimating a Bernoulli Parameter

Suppose $X_1, \ldots, X_n$ are i.i.d. samples from a Bernoulli distribution with unknown parameter $\theta \in [0, 1]$, where:

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x}, \quad x \in \{0, 1\}.$$

Let the observed data be $x_1, \ldots, x_n$, with $S = \sum_{i=1}^{n} x_i$ denoting the number of successes.
Then the likelihood function is:

$$L(\theta) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^S (1 - \theta)^{n-S}.$$

The log-likelihood is:

$$\ell(\theta) = S \log \theta + (n - S) \log(1 - \theta).$$

To find the maximum, we differentiate and set to zero:

$$\frac{d\ell}{d\theta} = \frac{S}{\theta} - \frac{n - S}{1 - \theta} = 0.$$

Solving:

$$\frac{S}{\theta} = \frac{n - S}{1 - \theta} \quad \Rightarrow \quad \hat{\theta}_{\text{MLE}} = \frac{S}{n}.$$

Thus, the MLE of θ is the sample mean: the proportion of successes. This simple example highlights the natural interpretation of MLE: it finds the parameter that best matches the empirical data.

Asymptotic Properties

Under regularity conditions, the MLE enjoys desirable asymptotic properties:

- **Consistency:** $\hat{\theta}_{\text{MLE}} \to \theta$ in probability as $n \to \infty$.
- **Asymptotic Normality:** $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$, where $I(\theta)$ is the Fisher information.
- **Asymptotic Efficiency:** The MLE achieves the Cramér-Rao lower bound asymptotically.

Conclusion

Maximum Likelihood Estimation offers a principled and widely applicable method for estimating parameters in statistical models. It connects theory with practice: given data and a model, MLE identifies the parameter values that render the observed data most plausible. Its generality, consistency, and strong asymptotic properties make it a cornerstone of classical statistics and a workhorse of modern machine learning.

Historical Context: Maximum Likelihood Estimation was introduced by the statistician Ronald Fisher in the early 20th century. Fisher's key insight was that many statistical problems can be solved by choosing the parameters of a model that make the observed data most probable. This approach unified estimation methods and became one of the most fundamental tools in statistics. MLE connects well with probability theory and has wide applications, from genetics to machine learning.

Why Maximum Likelihood? In statistics, we often have data generated by some unknown process described by parameters. The goal of MLE is to find the parameter values that best explain the observed data. By defining a likelihood function (the probability of observing the data given parameters), MLE picks the parameters that maximize this function, thus providing the most "likely" explanation.

Applications of MLE: MLE is widely used because it produces estimates with good theoretical properties: under mild conditions, MLE estimators are consistent (they get closer to the true value as data grows) and efficient (they have the smallest possible variance among unbiased estimators). It forms the backbone of many models and is implemented in most statistical software.

Maximum Likelihood Estimation (MLE) is a cornerstone of modern statistical inference. Developed in the early 20th century by Sir Ronald A. Fisher, MLE provides a systematic framework for estimating the parameters of a probabilistic model. Fisher introduced the method in the 1920s, formalizing it as a rigorous alternative to the method of moments and laying the groundwork for much of classical and modern statistical theory. MLE has since become one of the most widely used estimation techniques due to its generality, mathematical tractability, and strong theoretical properties. It applies to a broad class of models, including both discrete and continuous distributions, and serves as the basis for many advanced statistical methods, including Generalized Linear Models (GLMs), Bayesian inference (as the likelihood term), and machine learning algorithms.

3.5.1 Motivation and intuition

MLE seeks the parameter θ that makes the observed data most probable under the assumed model. In other words, it chooses the parameter that maximizes the likelihood function:

$$L(\theta) = P(X_1 = x_1, \ldots, X_n = x_n \mid \theta)$$

Example: For a Bernoulli distribution with unknown probability p, the likelihood of observing a sequence of 0s and 1s is:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

3.5.2 The Likelihood and Log-Likelihood functions

For independent and identically distributed data $X_1, \ldots, X_n$ with density or mass function $f(x;\theta)$, the likelihood function is:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

To simplify differentiation, we often use the **log-likelihood**:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

To find the MLE $\hat{\theta}$:

1. Write down the log-likelihood $\ell(\theta)$.

2. Take the derivative with respect to θ: $\frac{d\ell}{d\theta}$.
3. Solve $\frac{d\ell}{d\theta} = 0$ to find critical points.
4. Check which value maximizes the likelihood (often via the second derivative or boundary checks).

Example: Bernoulli MLE
For $X_i \sim \text{Bernoulli}(p)$,

$$\ell(p) = \sum_{i=1}^{n} [x_i \log(p) + (1 - x_i) \log(1 - p)]$$

Taking derivative:

$$\frac{d\ell}{dp} = \sum_{i=1}^{n} \left[\frac{x_i}{p} - \frac{1 - x_i}{1 - p} \right] = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1 - p}$$

Solving yields:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Properties of the MLE
The importance of MLE lies not only in its general applicability but also in its powerful theoretical properties. Under regularity conditions, MLEs are asymptotically optimal estimators in the sense that they:

- **Consistency:** $\hat{\theta} \to \theta$ as $n \to \infty$
- **Asymptotic Normality:** $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$, where $I(\theta)$ is the Fisher information
- **Efficiency:** Asymptotically achieves the Cram'er-Rao lower bound
- **Invariance:** If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$ for any differentiable function g

These properties make MLE a preferred method in both theoretical and applied statistics, especially for large-sample inference. In practice, these properties justify the use of MLE even when exact finite-sample distributions are hard to derive.

Application to Generalized Linear Models
Generalized Linear Models (GLMs) are an important class of models that extend linear regression to non-normal response variables by using a link function and a distribution from the exponential family. MLE plays a central role in fitting GLMs because the estimation of the model parameters is achieved by maximizing the likelihood of the observed responses. For instance, in logistic regression—used for binary outcomes—the log-odds of success is modelled as a linear combination of predictors:

Example: Logistic Regression
For binary response data, logistic regression models the log-odds as a linear function of predictors:

$$\log\left(\frac{p}{1 - p}\right) = \beta_0 + \beta_1 x$$

MLE is used to estimate the coefficients β_0, β_1 by maximizing the binomial log-likelihood.

Exercises

1. Exercise [...].
2. Exercise [...].
3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Chapter 4

Introduction to hypothesis testing

The object of statistical science is the reduction of data to relevant information.
— Ronald A. Fisher

4.1 Prediction vs inference

In the previous chapters we have introduced the mathematical theory of probability. That is, we have developed a series of tools, a *theory*, which enables us to make predictions about stochastic processes. But, contrary to what is normally explained in introductory courses, science is not always headed in the *mathematical prediction followed by experimental measurement* direction. There are many cases, as we will soon see, where expectations - or *hypothesis* are formulated for some given phenomena, and no prediction is made. In such cases, where we focus on interpretability rather than forecast, it is from measurement that we will try to see, or *infer* if our hypothesis are compatible with given data. Indeed, most modern data analysis and hypothesis testing lie in the inferential statistics, rather than predictive probability [47], [48].

The philosophical division between prediction and inference has deep roots in the evolution of scientific reasoning. On the one hand we find prediction, which is more agnostic to explanation and utilitarian, ultimately concerned with *what will happen next*. Inference, in contrast, is rooted in classical traditions of epistemology and induction, and seeks to illuminate the underlying structure or cause of observed data. It asks, fundamentally, *why we see what we do*, invoking theories, latent mechanisms, and parameters that might explain the observed phenomena.

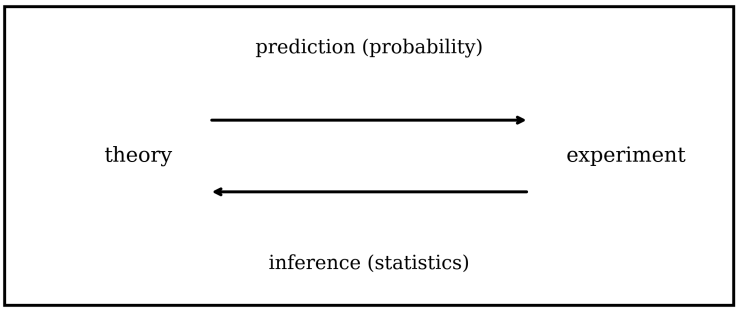


Figure 4.1: Representation of the predictive (from theory to experimental verification) and inferential (from data to underlying truth, descriptive and hypothesis testing) approaches to probability and statistics [49].

Consider the examples we saw in previous chapters: there we developed a series of mathematical tools, algorithmic-wise, that enabled us to make predictions for the so-called Bernoulli, Binomial, Gaussian, ..., distributed events. For a random variable such as the face of the coin, or dice, or any other stochastic event we could measure in a laboratory, we built *probability as a theory to predict* what would happen next, without going much deeper on why a certain set of events are Binomial, or Poisson, or Gaussian distributed, or why they follow these specific mathematical expressions.

This distinction becomes particularly important in statistical modeling. Predictive models, such as those deployed in machine learning, are often evaluated by their performance on unseen data, with little regard for the interpretability of parameters. Inference, however, thrives on interpretable structure — effect sizes, confidence intervals, causal assumptions — offering insights that extend beyond mere forecasting. Yet the two are not mutually exclusive: many contemporary methods, especially in Bayesian statistics, attempt to harmonize prediction and inference within unified frameworks. Still, when decisions hinge on understanding causation or evaluating mechanisms—as in medicine, economics, or the social sciences—inference retains a privileged role. We will review this topic in Chapter 6, when we focus on the Bayesian and frequentist definition of probability. Approaching modern times, the rise of probabilistic models in the 20^{th} century has made now possible to quantify uncertainty in both directions, but their goals remain distinct. Inference aims to learn about the data-generated process, while prediction aims to accurately forecast future observations, regardless - at times - of the process's inner interpretability [47]. [49].

Within this context, hypothesis testing emerges as a formal tool to carry out inference under uncertainty. Popularized in the early 20^{th} century by Karl Pearson and Ronald A. Fisher, and later formalized by Jerzy Neyman and Egon Pearson, hypothesis testing provides a structured way to assess whether observed data is consistent with a particular theoretical claim. At its core, and as we will see in the next section, the method tests a "null" hypothesis (typically representing chance or no effect) against an "alternative" (commonly representing an unexpected event). Given these assumptions, we will build *estimator* quantities from data, such as sample means and variances, and combine them to compute *informative quantities*, sometimes referred as *statistics* or *statistic tests*. Then, we will calculate the probability of observing some data as extreme as ours under the null expectation. While often misused or misinterpreted, hypothesis testing remains foundational in the sciences, offering a bridge between data and theory, between the probabilistic world of prediction and the explanatory realm of inference [50].

[Lindley] Inference

The formulation that has served statistics well throughout this century is based on the data having, for each value of a parameter, a probability distribution. This accords with the idea that the uncertainty in the data needs to be described probabilistically. It is desired to learn something about the parameter from the data. Generally not every aspect of the parameter is of interest, so write it as (Θ, α) where we wish to learn about Θ with α as a nuisance, to use the technical term [...].

4.2 Hypothesis, significance, p-values

The term *hypothesis testing* is normally used to refer a broad set of tools addressing parameter estimation, inference, and various exploratory analysis on random measurements and observations. In the last decades, expressions like hypothesis testing, hypothesis test, statistical inference - sometimes referred to as exploratory analysis - have gained popularity and become one of the standards in most experimental sciences, given the automatization of experiments and the large amounts of data available.

Once we have covered the idea of parameter estimation, sample distributions, and the idea of estimators, we will now formulate hypothesis on the *population*, or true - *unkwnown* - parameters, and then build *statistic tests* to quantify how far - or close - are these hypothesized values from the *sample*, observed, experimental, values. And finally, we will quantify how certain we are about the values obtained - how *significant* they are - computing the *p-value*, standing from Pearson value [...].

The general approach we will follow, regardless of the kind of question we are after and the observations made, can be summarized as follows:

- Formulate *null* hypothesis H_0 and *alternative* hypothesis H_1 about the *true - population* parameters, generally for the mean or variance, *prior to experiment*.
- Collect data, make observations, make measurements.
- Compute *estimators* and *informative quantities* from our observed values, normally referred to as *statistics*, or *statistic tests*.
- Compute p-value, the probability that *we obtained a value at least as extreme as the one we obtained for our statistic test*.
- Accept or reject the null hypothesis, based on the p-value, which quantifies how probable was to see this result.

Here we should pause for some time, and we must review some basic concepts about integration and probabilities, as we will try to build a *mathematically accurate definition* of significance and the idea of p-value. The idea of p-value was developed by Pearson in the 1920s - indeed, p comes from no other than Pearson-value, and it was built already on the framework of estimator quantities and statistic tests.

Recall for a moment the idea of cumulative probability. Take some random variable with known distribution (e.g., the probability of obtaining 3 times heads in 10 tosses of a coin, which we know follows a Binomial distribution).

$$P(x = x_0; n, k) = \binom{n}{x_0} p^x (1-p)^{n-x}, \tag{4.1}$$

And ask now what is the probability of obtaining *more than* 3 times heads. We know that this is given by the cumulative probability,

$$P(x > x_0; n, k) = \sum_{x=x_0}^{x_n} \binom{n}{x_i} p^x (1-p)^{n-x}, \tag{4.2}$$

And the analogous in a continuous variable

$$f(x > x_0) = \int_{x_0}^{\infty} f(x)dx \; . \tag{4.3}$$

Now, with these concepts in mind, let's imagine a general, simple hypothesis testing scenario. We have some random variable x following a gaussian distribution, and a hypothesized value μ for its true - or *population* mean. As we saw in Chapter 3, we can make a set of measurements, and out of these compute an estimator, for instance, a sample mean $\bar{x}$. The real hypothesis testing problem is now to compare these two quantities, and somehow *quantify* how similar - or different - they are. There are many ways of doing this, but one simple case is to compute just a difference between the two.

$$t = \frac{\bar{x} - \mu}{\sigma} \quad . \tag{4.4}$$

This is normally called a t statistic, or a t *statistic* test, which we will describe in detail in the next section. For now, just consider it a useful quantity representing how similar the true mean and the sample mean are to each other. Note that as μ tends to $\bar{x}$, the t variable tends to zero.

It is now, once we have computed a statistic, an informative quantity from our data, that we can ask: what was the probability, given that first assumption about the true mean, that we obtain this particular result for the t variable? Note that the t variable is computed out of our estimator $\bar{x}$, which was computed out of our random observations. Hence, the t variable, and any other statistic test we could have built - a

Fisher, a $\chi 2$, etc - *is a random variable as well*, and it will follow *some* distribution.

Now, from our set of observations, we will obtain a specific value, t_{obs}, standing for *observed* t. Then, the p-value is defined as the probability of obtaining a value *at least as extreme* as the one we obtained for our statistic, given our hypothesized value for the true - or population - parameter. In other words, given the null hypothesis was true. The *at least as extreme* part is what enables the cumulative distribution.

$$p = P(t > t_{obs}) = \int_{t_{obs}}^{\infty} f(t)\, dt, \tag{4.5}$$

For the computation of the numerical value we just need to know what is the probability density of the t variable. This way, we are just left with solving a simple integral [...].

4.3 Statistical tests: some examples

4.3.1 Compare sample mean with hypothesized value - One sample t-test

The t-test, a cornerstone of inferential statistics, emerged from practical necessity in the early 20th century, in no other than the world of brewing: William Sealy Gosset, a statistician employed by the Guinness Brewery in Dublin, devised the method to address the problem of making reliable inferences from small sample sizes - a common challenge in quality control. Because Guinness forbade its employees from publishing research, Gosset adopted the pseudonym "Student," and the t-test entered the statistical canon as "Student's t-test." Far from being a mere curiosity, this test helped lay the foundation for modern statistical thinking, offering a systematic way to evaluate whether observed differences could be attributed to chance [46].

Contextually, the t-test emerged during a period of growing interest in the mathematical modeling of uncertainty. While the 19th century had seen the establishment of probability theory and the development of the normal distribution by figures like Gauss and Laplace, it was not until the early 20th century that statisticians began to address the problem of small samples and experimental variability. Gosset's innovation came at a time when the rigour of scientific inquiry was undergoing a deep transformation; where laboratory science, agricultural experimentation, and industrial quality control all demanded more precise methods of inference. The t distribution, with its tails heavier than the normal distribution, conveniently modelled the added uncertainty inherent in small datasets, capturing the low confidence in inferences drawn from limited observations.

In the broader sweep of statistical development, the t-test exemplifies the bridging of theoretical elegance and empirical utility, making it feasible for practitioners across disciplines to assess hypotheses without the need for large-scale experimentation. Its adaptability, whether in comparing means between two groups or assessing paired observations, has ensured its enduring presence in the researcher's toolkit. Over a century later, the t-test remains not merely a relic of an industrial past but a living method, still invoked in clinical trials, psychological studies, and countless scientific explorations, bearing with it both the legacy of its creator and the evolving rigour of modern analysis.

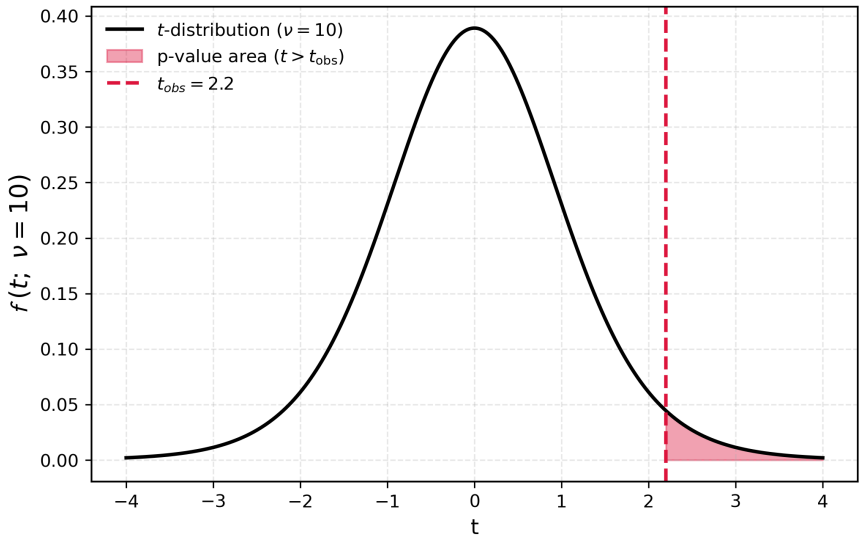


Figure 4.2: Representation of the t statistic, following the Student's t distribution, for a particular value of the degrees of freedom ($\nu = 10$). The integral of the shadowed area represents the *1-sided*, or *1-tailed* p-value, as the probability of obtaining a result *at least as extreme* as the one obtained t_{obs}.

The student's t is used to compare the sample mean $\bar{x}$ of a set of observations to ha hypothesized value μ.

It assumes that the sample data are drawn from a normally distributed population, hence it is an example of a *parametric* test. We will discuss more about parametric and non-parametric observations, and how to test for normality further in the chapter. The t *statistic test*, or t *statistic*, or just t *variable* for simplicity, is given by:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \tag{4.6}$$

where $\bar{x}$ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size. As we se, it is just a measure of how far the sample mean $\bar{x}$ and the hypothesized true value μ are from each other. Indeed, it is built in such a way that, as the sample mean $\bar{x}$ gets closer to the hypothesized value μ, the t-variable approaches zero.

Then, given some data was observed and and we obtained a specific value for our t - let's call it t_{obs}, to compute a p-value we just need to compute what was the probability of that particular value. To do that, we just recall our t variable was indeed a random variable depending on our random observations, which produced some random sample mean and variance, and some degrees of freedom $n - 1$

$$p = P\left(t > tobs\right) = 2 \cdot \int_{|t|}^{\infty} f_{T_{n-1}}(x)\, dx = 2 \cdot \left[1 - F_{T_{n-1}}(|t|)\right] \tag{4.7}$$

The t distribution

Being $f_{T_{n-1}}$ the PDF of the t variable, the *Student's t distribution* with $n - 1$ degrees of freedom, and $F_{T_{n-1}}$ the corresponding cumulative distribution, as we discussed in chapter 2 [...]. Here, we are computing the probability of t being greater than the one we obtained, and we do that just by integrating the t-distribution [...]. Note that here we are computing a 2-sided p-value, hence te factor 2 at the beginning.

Given a sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, the one-sample t-statistic is defined as:

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

where $\bar{X}$ is the sample mean, S is the sample standard deviation, and μ_0 is the hypothesized population mean.

Under the null hypothesis $H_0 : \mu = \mu_0$, the statistic follows a Student's t-distribution with $n - 1$ degrees of freedom:

$$t \sim t_{n-1}$$

The probability density function (PDF) of the t-distribution with ν degrees of freedom is:

$$f(t;\ \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{4.8}$$

Distribution and computation of the p-value as defined above.

4.3.2 Compare sample means of two independent groups - Two sample t-test

The two-sample t-test represents a natural extension of Gosset's original formulation, expanding its reach from single-sample inference to the comparison of two independent groups. This test seeks to determine whether the sample means, $\bar{x}_1$ and $\bar{x}_2$ of two populations are significantly different. It operates under the assumption that the underlying distributions are approximately normal and that the variances between groups are equal or at least comparable, though variants like Welch's t-test have relaxed these conditions. The test statistic itself balances the observed difference in means against the pooled standard error, weighing signal against noise [50].

The historical roots of the two-sample t-test are intertwined with the rise of controlled experimentation in the biological and social sciences during the early to mid-20th century. Ronald Fisher, Jerzy Neyman, and Egon Pearson contributed to formalizing the logic of hypothesis testing, and the t-test became a practical tool within this growing paradigm. Its popularity grew not merely because it was mathematically sound, but because it was accessible. A concise, simple and interpretable method for testing simple but essential questions in science: does treatment differ from control, is an observed effect more than a fluke, etc.

Even in today's data-rich landscape, the two-sample t-test remains remarkably resilient. Its conceptual clarity and computational simplicity make it a first recourse in evaluating differences between two conditions, from clinical trials comparing drug efficacies to educational studies assessing learning interventions. In many ways, it is the distilled essence of statistical thinking: discerning pattern from randomness, while acknowledging uncertainty. Its elegance lies not in its complexity, but in its ability to render the ordinary rigorously meaningful.

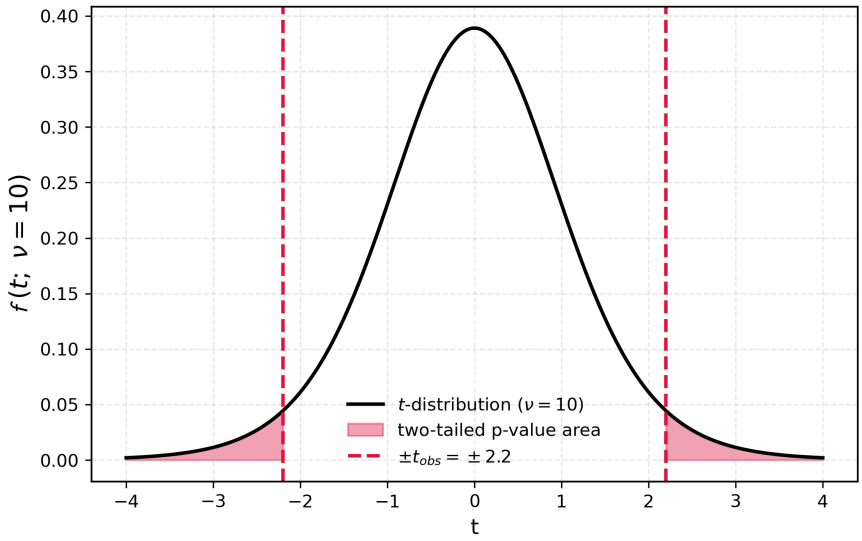


Figure 4.3: Representation of the t statistic, following the Student's t distribution, for a particular value of the degrees of freedom ($\nu = 10$). The integral of the shadowed area represents the *2-sided*, or *2-tailed* p-value, as the probability of obtaining a result *at least as extreme* as the one obtained t_{obs}.

The so-called *two-sample t-test* is used to determine whether the sample means $\bar{x}_1$ and $\bar{x}_2$ of two sets of observations are significantly different from one another. It assumes that the sample data are drawn from a normally distributed population, hence it remains an example of parametric test. The t *statistic test*, or t *statistic*, or just t *variable* for simplicity, is given by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2(n_1 - 1)^2 + s_2^2(n_2 - 1)^2}}, \tag{4.9}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, s_1^2 and s_2^2 are the sample variances, n_1 and n_2 the sample sizes.

The computation of the p-value:

$$p = P\left(t > tobs\right) = 2 \cdot \int_{|t|}^{\infty} f_{T_{df}}(x)\, dx = 2 \cdot \left[1 - F_{T_{df}}(|t|)\right] \tag{4.10}$$

Being $f_{T_{n-1}}$ the PDF of the t variable, the *Student's t distribution* with $n_1 + n_2 - 1$ degrees of freedom, and $F_{T_{n-1}}$ the corresponding cumulative distribution Note here we assume equal variances. For Welch's t-test (unequal variances), use the same form, but with Welch-adjusted [...]

The t distribution

For two independent samples $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$, the test statistic is:

$$t = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

with pooled variance estimate:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n + m - 2}$$

Under $H_0 : \mu_1 = \mu_2$, the statistic follows a t-distribution with $n + m - 2$ degrees of freedom:

$$t \sim t_{n+m-2}$$

Distribution and computation of the p-value as defined above.

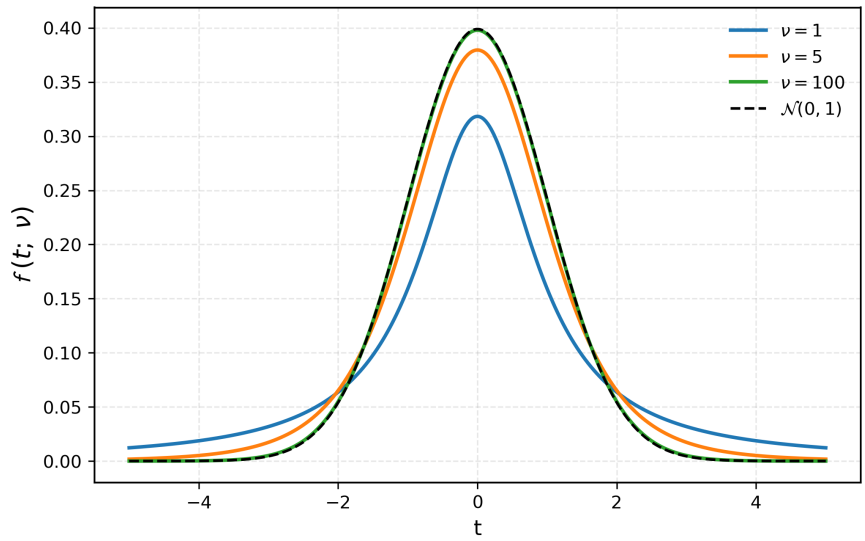


Figure 4.4: Representation of the Student's t distribution, for different values of the degrees of freedom ν.

4.3.3 Compare sample variances of two groups - Fisher's exact test

Fisher's exact test, introduced by Ronald A. Fisher in the 1930s, occupies a unique place in the statistical repertoire as a method built not on approximation, but on exact combinatorial reasoning. Designed to test for nonrandom association between two categorical variables in a 2x2 contingency table, the test calculates the exact probability of obtaining the observed configuration—or one more extreme—under the null hypothesis of independence. Unlike the chi-squared test, which relies on asymptotic assumptions, Fisher's exact test makes no concession to large-sample theory, rendering it particularly well-suited for small datasets where expected counts may be sparse [50].

The genesis of the test is as intellectual as it is practical. Fisher, with his characteristic synthesis of biological intuition and statistical rigor, crafted the test to address problems in genetics and experimental design. In particular, he sought a method that preserved the integrity of inference even when data were limited—a not uncommon situation in biological sciences of his time. His test exemplifies his broader philosophical stance: that statistical methods should be exact, and that inferences should reflect the full structure of the data, not rely on approximations that may distort conclusions.

In modern usage, Fisher's exact test continues to shine in small-sample contexts—such as clinical trials, epidemiology, and case-control studies—where cell counts are low and the cost of misjudging association is high. Its continued relevance, despite the availability of powerful computational tools and large datasets, underscores a broader truth: that precision in reasoning often matters as much as the scale of inquiry. The test embodies an ideal of statistical craftsmanship, where fidelity to the data is preserved down to the last permutation.

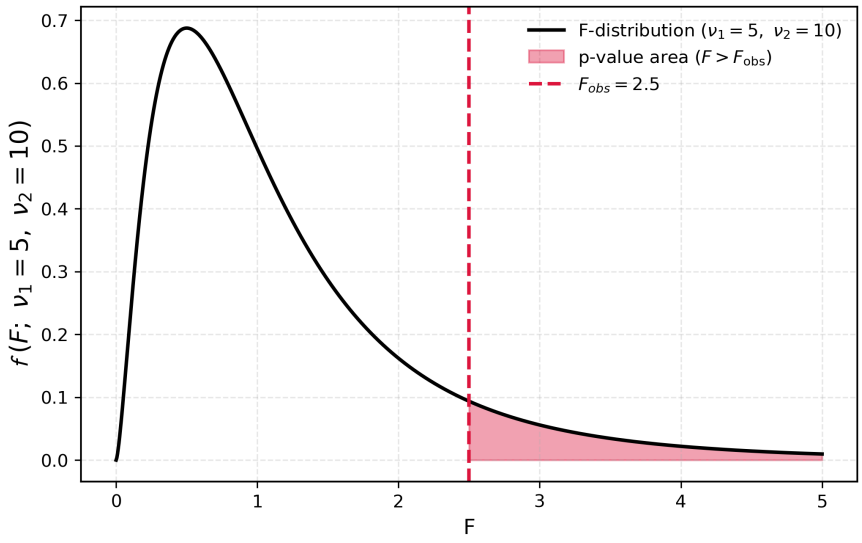


Figure 4.5: Representation of the F statistic, following the Fisher distribution, for a particular value of the degrees of freedom ($\nu = 10$). The integral of the shadowed area represents the *1-sided*, or *1-tailed* p-value, as the probability of obtaining a result *at least as extreme* as the one obtained t_{obs}.

The next example we will encounter is an extension of the same question., The so-called *Fisher t-test*, or just F test, is used to determine whether the sample variances of two sets of observations are significantly different from one another. It assumes that the sample data are drawn from a normally distributed population.

The F statistic is a ratio of two independent variance estimates, each scaled by their respective degrees of freedom. It is used to test whether group variances (or group means, in ANOVA) differ significantly. The

general form of the F statistic is:

$$F = \frac{S_1^2/\nu_1}{S_2^2/\nu_2}$$

where s_1^2 and s_2^2 are the sample variances, and the degrees of freedom are $d_1 = n_1 - 1$ and $d_2 = n_2 - 1$.

The computation of the p-value:

$$p = \sum_{\text{extreme values}} \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

Under the null hypothesis, the F statistic follows the F-distribution:

$$F \sim F(\nu_1, \nu_2)$$

and the p-value is computed as the upper-tail probability:

$$p = P(F_{\nu_1,\nu_2} \geq F_{\text{obs}}) = \int_{F_{\text{obs}}}^{\infty} f_{F_{\nu_1,\nu_2}}(x)\, dx$$

The F distribution

When the assumption of equal variances is violated, the test statistic is:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

with an approximate degrees of freedom:

$$\nu \approx \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{(S_X^2/n)^2}{n-1} + \frac{(S_Y^2/m)^2}{m-1}}$$

Under the null $H_0 : \mu_1 = \mu_2$, the distribution of t is approximated by:

$$t \sim t_\nu$$

Distribution and computation of the p-value as defined above.

4.3.4 Compare variation on more than two groups - Fisher's ANOVA

The analysis of variance, or ANOVA, is one of the 20th century's major statistical innovations, largely credited to Ronald Fisher, who formalized it in the 1920s while working at the Rothamsted Experimental Station. Its core purpose is to partition total variation in a dataset into components attributable to different sources—typically, between-group variation and within-group (error) variation. ANOVA provides a systematic method for comparing more than two means simultaneously, resolving what would otherwise require a series of pairwise t tests and an increasing risk of Type I error [47].

Conceptually, ANOVA represents a unifying idea: that observed data contain structure and randomness, and that the task of statistical inference is to disentangle them. It was particularly well-suited to agricultural experiments, where multiple treatments were applied across randomized plots, and variation needed to be measured with mathematical elegance and practical clarity. Fisher's framework gave researchers a way to understand experimental results in terms of their statistical "signal," paving the way for rigorous design and interpretation across many scientific disciplines.

Today, ANOVA remains foundational in experimental science, psychology, and the social sciences. Its principles have expanded into more complex models—repeated measures, factorial designs, and mixed effects frameworks—but the essential logic endures. It is a statistical lens through which variability is not merely noise, but an object of structured inquiry. In this way, ANOVA reflects a deeper epistemological commitment: that difference can be measured, that structure can be inferred, and that complexity, when properly modeled, reveals its underlying form.

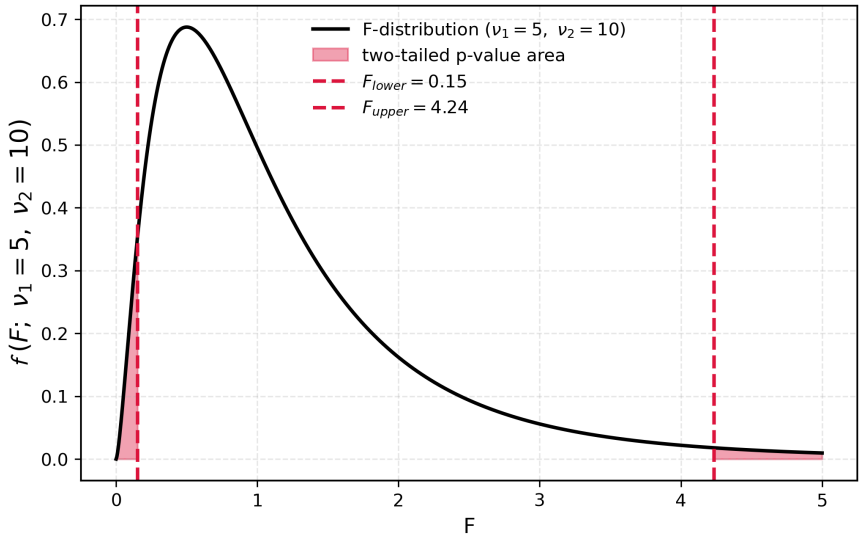


Figure 4.6: Representation of the F statistic, following the Fisher distribution, for a particular value of the degrees of freedom ($\nu = 10$). The integral of the shadowed area represents the *2-sided*, or *2-tailed* p-value, as the probability of obtaining a result *at least as extreme* as the one obtained t_{obs}.

The so-called Analysis of Variance, one way ANOVA, or just ANOVA, is used to determine whether the variation of a dataset comes primary from variation within the samples themselves, or from variation between the groups. It is an extension of the Fisher test, where the F statistic is computed as:

$$f(x; d_1, d_2) = \frac{s^2_{\text{between}}}{s^2_{\text{within}}},$$

where s_1^2 and s_2^2 are the sample variances, and the degrees of freedom are $d_1 = n_1 - 1$ and $d_2 = n_2 - 1$.

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{SS_{\text{between}}/(k-1)}{SS_{\text{within}}/(N-k)}$$

where:

- SS_{between} is the sum of squares between groups,
- SS_{within} is the sum of squares within groups,
- k is the number of groups,
- N is the total number of observations.

The computation of the p-value:

$$p = \int_F^\infty f_{F_{df_1,df_2}}(x)\,dx = 1 - F_{F_{df_1,df_2}}(F)$$

The F distribution

Assume k groups with sample sizes n_i and group means $\bar{X}_i$. The test statistic is:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{SS_{\text{between}}/(k-1)}{SS_{\text{within}}/(N-k)}$$

with:

$$SS_{\text{between}} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2, \quad SS_{\text{within}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

Under $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, the statistic follows:

$$F \sim \mathcal{F}_{k-1,N-k}$$

The PDF of the $\mathcal{F}_{d_1,d_2}$ distribution is:

$$f(x) = \frac{1}{\mathrm{B}(d_1/2, d_2/2)} \left(\frac{d_1}{d_2}\right)^{d_1/2} \frac{x^{d_1/2-1}}{\left(1 + \frac{d_1}{d_2}x\right)^{(d_1+d_2)/2}}$$

for $x > 0$.

Distribution and computation of the p-value as defined above.

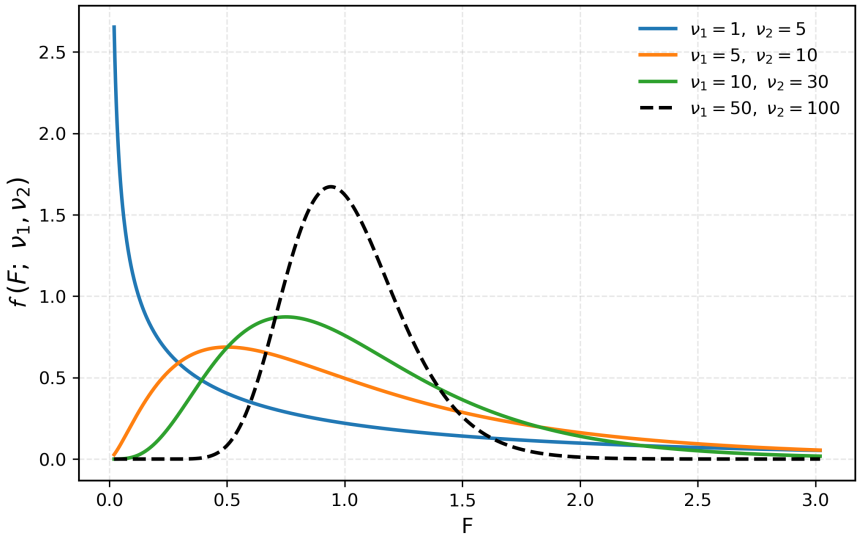


Figure 4.7: Representation of the Fisher distribution, for different values of the degrees of freedom ν_1, ν_2.

4.3.5 Compare distributions and testing for normality - χ^2 test

The chi-squared test, with its roots in the work of Karl Pearson at the turn of the 20th century, stands as one of the earliest formal tests of goodness-of-fit and independence. Built on the comparison of observed frequencies to expected ones under a given hypothesis, the chi-squared statistic accumulates discrepancies between what is seen and what is statistically anticipated. It is asymptotic in nature, relying on the approximation of the chi-squared distribution, which becomes increasingly accurate with larger sample sizes and expected counts [49].

Pearson's motivation was both mathematical and empirical: to provide a quantitative measure for evaluating how well a theoretical model matched observed data, particularly in biological contexts such as Mendelian genetics. His test became a cornerstone of categorical data analysis, offering a simple yet powerful method for detecting associations in contingency tables and deviations from expected distributions. Over time, the test was expanded and refined, notably by Fisher and Yates, but it has retained its essential character as a test of fit between expectation and observation.

In the present day, the chi-squared test continues to serve as an indispensable tool across fields as varied as market research, epidemiology, and political science. Its ability to handle complex tables and large datasets makes it well-suited to the modern data deluge. Yet its appeal also lies in its conceptual clarity: the squared discrepancy between what the world appears to be and what a hypothesis claims it should be. It is a statistical mirror, held up to the categorical world, revealing the tensions between model and reality with a cold but elegant precision.

The so-called Pearson χ^2-test is used to determine whether the set of observations is significantly different from some expected - or hypothesized - values. The χ^2 statistic is given by:

$$\chi^2(x; N-1) = \sum_{i=1}^{N} \frac{(O_i - Ei)^2}{E_i},$$

It is also used to test for normality and evaluate the goodness of a fit [...].

The computation of the p-value:

$$p = \int_{\chi^2}^{\infty} f_{\chi^2_{df}}(x)\, dx = 1 - F_{\chi^2_{df}}(\chi^2)$$

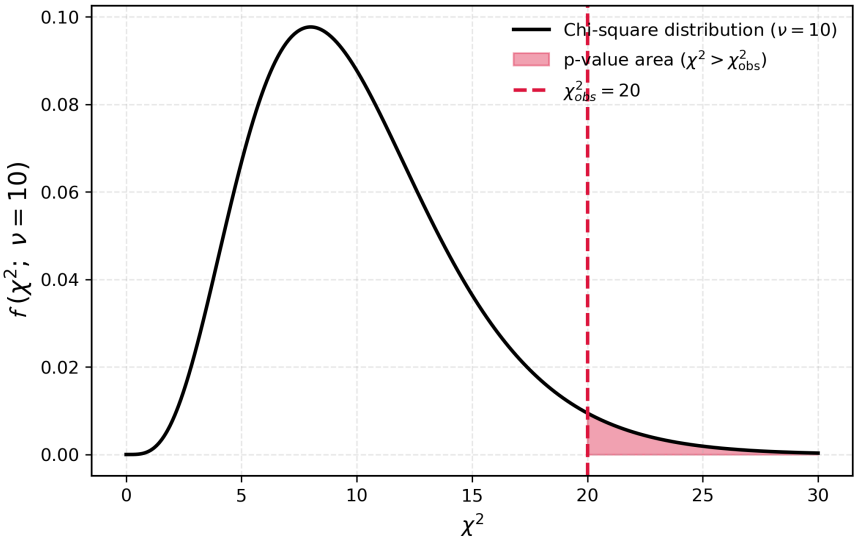


Figure 4.8: Representation of the chi^2 statistic, following the Pearson chi^2 distribution, for a particular value of the degrees of freedom ($\nu = 10$). The integral of the shadowed area represents the *1-sided*, or *1-tailed* p-value, as the probability of obtaining a result *at least as extreme* as the one obtained chi^2_{obs}.

The χ^2 distribution

For observed O_{ij} and expected counts E_{ij} in an $r \times c$ table, the test statistic is:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Under the null hypothesis of independence, the distribution is:

$$\chi^2 \sim \chi^2_{(r-1)(c-1)}$$

The PDF of the chi-squared distribution with k degrees of freedom is:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x > 0$$

Distribution and computation of the p-value as defined above.

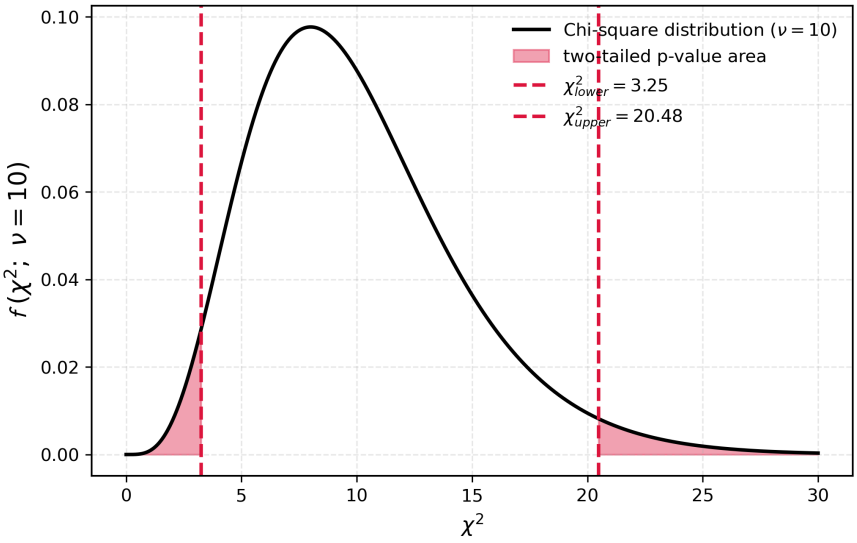


Figure 4.9: Representation of the chi^2 statistic, following the Pearson chi^2 distribution, for a particular value of the degrees of freedom ($\nu = 10$). The integral of the shadowed area represents the *2-sided*, or *2-tailed* p-value, as the probability of obtaining a result *at least as extreme* as the one obtained chi^2_{obs}.

4.4 Parametric and non-parametric

Parametric and non-parametric

1. **Student's *t*-Test (One-Sample and Two-Sample)**
   Introduced by William S. Gosset under the pseudonym "Student" in his seminal paper [46], this test allows inference on population means when the standard deviation is unknown and the sample size is small. The two-sample version compares the means of two independent groups under assumptions of normality and equal variances.
   Non-parametric alternative: The **Mann–Whitney U test** (also called the Wilcoxon rank-sum test) for independent samples, and the **Wilcoxon signed-rank test** for paired or single-sample symmetry testing.

2. **Fisher's Exact Test**
   Formulated by R. A. Fisher in his 1935 treatise on experimental design [48], this exact test computes the hypergeometric probability of a given 2x2 contingency table under the null hypothesis of independence. It is particularly suited for small samples, where asymptotic methods like chi-squared tests may not be valid.
   Non-parametric alternative: The test is already exact and non-parametric; however, **Barnard's exact test** is sometimes proposed as a more powerful alternative in small-sample settings.

3. **Analysis of Variance (ANOVA)**
   Fisher also introduced ANOVA in his earlier work [47], laying the foundation for modern experimental statistics. ANOVA decomposes total variation into between-group and within-group components, enabling testing of the equality of multiple means.
   Non-parametric alternative: The **Kruskal–Wallis H test**, which operates on ranks rather than means, is used when ANOVA assumptions (e.g., normality or homoscedasticity) are violated.

4. **Chi-Squared Test**
   Originally proposed by Karl Pearson in 1900 [49], the chi-squared test evaluates how observed frequencies compare to expected frequencies under a null model of independence or goodness-of-fit. It is widely used in contingency table analysis and categorical data testing.

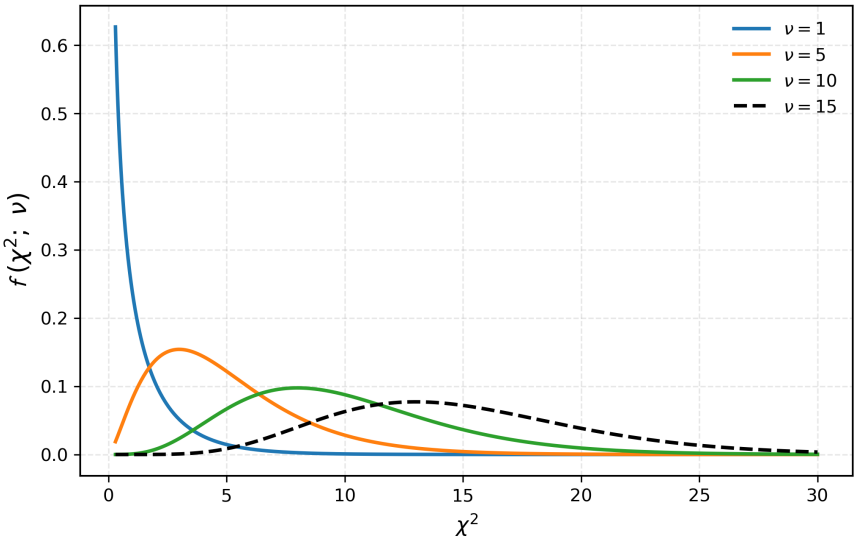


Figure 4.10: Representation of the *chi*^{2} distribution, for a particular value of the degrees of freedom ($\nu = 10$).

Non-parametric alternative: While the chi-squared test is non-parametric in principle, **Fisher's exact test** is often preferred for 2x2 tables with small expected counts.

5. **Welch's *t*-Test (Unequal Variance Version)**
   To address the limitations of assuming equal variances in the classical *t*-test, Welch proposed a generalization in 1947 [50] that remains robust under heteroscedasticity. It uses a weighted estimate of variance and adjusts the degrees of freedom accordingly.
   Non-parametric alternative: The **Mann–Whitney U test** again serves as a distribution-free method for comparing central tendencies when variance assumptions fail.

4.5 Comparing data and normalization

Comparing data and normalization

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Chapter 5

Introduction to bayesian probability

Probability statements are just summaries of repeated observations.

— W. V. Quine

5.1 Motivation and philosophy

Human reasoning has always been entangled with uncertainty. Whether deciding whether dark clouds herald rain, or whether a patient's symptoms imply disease, we rely on judgments that are conditional on partial evidence. The language of probability emerged to give structure to this uncertainty. In particular, conditional probability provides a disciplined way of asking: how does the plausibility of one event shift when another is known to occur? Without such a concept, reasoning under uncertainty would remain vague, prone to bias, and resistant to mathematical analysis [54].

The Bayesian perspective takes this idea further, treating probability as a measure of belief that can be updated in light of new evidence. Originating with Thomas Bayes in the 18th century [55], this approach formalized how one ought to revise expectations when confronted with data. Pierre-Simon Laplace expanded Bayes' insight into a universal method of reasoning, claiming that "probability is common sense reduced to calculation" [56]. This philosophical orientation views uncertainty not only as randomness in the world, but as incompleteness in our knowledge.

From its inception, probability has had two competing interpretations: the frequentist view, which grounds probability in long-run relative frequencies of repeated events, and the Bayesian view, which interprets probability as rational degrees of belief. This tension has fueled centuries of debate. Yet, whichever interpretation one adopts, conditional probability and Bayes' theorem remain central: both perspectives agree that information alters the structure of uncertainty. To study these concepts is not merely to learn a technique, but to touch upon the foundations of rational inference itself.

5.2 Foundations of conditional probability

At its simplest, conditional probability quantifies how the likelihood of an event changes once we restrict our attention to a smaller world of possibilities. If A and B are events with $P(B) > 0$, the conditional probability of A given B is defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

This definition encapsulates the intuition that to assess A under the knowledge of B, one rescales the probability measure to the reduced universe where B is certain.

Independence of events is naturally expressed through this lens. Two events A and B are independent precisely when

$$P(A \mid B) = P(A),$$

that is, knowledge of B alters nothing about the chance of A. Conversely, when $P(A \mid B) \neq P(A)$, we have dependence, the essential source of informational value in probabilistic reasoning. For example, in medical testing, the presence of a symptom significantly alters the probability of a disease, making the conditional perspective indispensable [57].

Two fundamental tools arise from the definition: the *law of total probability*, which expresses the probability of an event as a weighted sum over a partition of the sample space, and *Bayes' theorem*, which inverts conditional probabilities. Bayes' theorem states that for events A and B,

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

providing the precise mechanism by which evidence (B) reshapes our belief in a hypothesis (A). In applied contexts, this formula underlies diagnostic testing, spam detection, and machine learning classifiers.

5.3 Bayesian reasoning and applications

While Bayes' theorem can be presented as a mere algebraic identity, its deeper power lies in its interpretation. In Bayesian reasoning, probabilities represent not just frequencies, but subjective degrees of belief. Bayes' rule then formalizes the principle of learning: one begins with a prior probability $P(A)$, updates with the likelihood $P(B \mid A)$, and arrives at a posterior probability $P(A \mid B)$. Symbolically,

$$\text{Posterior} \;\propto\; \text{Likelihood} \times \text{Prior}.$$

This deceptively simple formula encodes the cycle of rational inference: hypotheses are weighted, tested against data, and refined [58].

Consider medical diagnostics. Suppose a disease has prevalence $P(D) = 0.01$, and a test has sensitivity $P(\text{Positive} \mid D) = 0.99$ and false positive rate $P(\text{Positive} \mid \neg D) = 0.05$. If a patient tests positive, the posterior probability of disease is

$$P(D \mid \text{Positive}) = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.05 \times 0.99} \approx 0.167.$$

Contrary to intuition, a positive test does not imply near certainty of disease; the rarity of the disease weighs heavily against it. Such examples illustrate the necessity of Bayesian reasoning to temper human judgment, which often underestimates the influence of base rates.

Beyond medicine, Bayesian inference permeates contemporary science and technology. From machine learning algorithms that classify emails as spam, to Bayesian networks that model causal dependencies, to posterior simulations that calibrate cosmological models, the Bayesian framework provides a unifying language of uncertainty. It offers not merely a computational trick, but a philosophy: knowledge grows not by abandoning prior beliefs, but by systematically updating them in light of new evidence.

5.4 Rigorous mathematical formalism

To study probability rigorously, one typically begins with the idea of a *probability space*. This consists of three parts: a sample space Ω of all possible outcomes, a collection $\mathcal{F}$ of events (subsets of Ω) that are considered measurable, and a probability measure P that assigns numbers between 0 and 1 to these events, with $P(\Omega) = 1$. This framework, developed systematically by Kolmogorov [**?**], ensures that probability rests on a consistent mathematical foundation.

Conditional probability, as defined earlier, fits neatly into this setting. When $P(B) > 0$, we define

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

This formula can be seen as constructing a new probability measure in which B is certain. Under this measure, the probabilities of other events are recalibrated accordingly.

Bayes' theorem can also be expressed in this language. If events $A_1, A_2, \ldots, A_n$ form a partition of Ω, then for any event B with $P(B) > 0$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)\, P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j) P(A_j)}.$$

This form emphasizes that posterior probabilities are simply normalized weights: the numerator favors hypotheses that make the evidence likely, while the denominator ensures the probabilities sum to one. This balance between rigor and intuition makes conditional probability accessible to students without advanced measure theory, while still preserving mathematical precision.

The rigorous formulation begins with the notion of *sample space*, denoted by Ω, introduced in the early 20th century by French mathematician Henri Lebesgue, among others. By *sample space* Ω, we will mean just the set of all possible outcomes of a certain measurement. The other key element is the *measure* μ, introduced in its modern form by Lebesgue in 1902 [38], and later adapted to probability as $\mathbb{P}$. Together they denoted an abstract mathematical object called the *measure space* (Ω, μ), which will eventually become the mathematical language used to described probability. If you want to read more about how measure theory led to probability theory in its its modern form, navigate through the works of Andrey Kolmogorov, specially his 1933 monograph *"Foundations of Theory of Probability* [26], where he formalized probability spaces using measure-theoretic language.

There have been at least two successful attempts to formalize probability, namely the Kolmogorov formulation and the Cox formulation. In Kolmogorov's formulation (see also probability space), sets are interpreted as events and probability as a measure on a class of sets. In Cox's theorem, probability is taken as a primitive (i.e., not further analyzed), and the emphasis is on constructing a consistent assignment of probability values to propositions. In both cases, the laws of probability are the same, except for technical details [...].

5.5 Stochasticity and Markov processes

So far, our discussion has treated probability as static, focused on isolated events and their interrelations. Yet in much of life, uncertainty unfolds dynamically over time: stock markets rise and fall, patients progress through stages of illness, and the weather changes from day to day. To capture such phenomena, probability theory turns to *stochastic processes*, families of random variables indexed by time. A stochastic process $\{X_t\}_{t \geq 0}$ represents the evolution of uncertainty itself [59].

Among these, the class of *Markov processes* is especially influential. A process is Markovian if its future depends only on its present state, not on its full history:

$$P(X_{t+1} \mid X_t, X_{t-1}, \ldots, X_0) = P(X_{t+1} \mid X_t).$$

This "memoryless" property is both a simplifying assumption and a powerful modeling principle: the present state contains all the relevant information from the past needed to predict the future.

A classic example is weather forecasting. Suppose tomorrow's weather depends only on today's, not on the sequence of previous days. Let the states be Sunny (S) and Rainy (R). We specify a transition matrix:

$$P = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix},$$

where the first row gives $P(\text{next} \mid S)$ and the second row gives $P(\text{next} \mid R)$. If today is sunny, there is an 80% chance tomorrow will also be sunny, and a 20% chance of rain. This simple model already captures persistence and variability in weather patterns.

The weather example illustrates both the limitations and the strengths of the Markov assumption. Of course, real weather depends on far more than just today's state; yet the Markov chain offers a tractable first approximation. More generally, Markov processes underpin statistical physics, genetics, economics, and machine learning, where complexity is tamed by assuming that the present state suffices to chart the course of uncertainty.

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Appendix A

Appendix: Vectors and matrices: a quick review

A.1 The roots and rise of algebra

Algebra and vectors form two essential branches of mathematics, each with a rich history and deep connections to the physical and abstract worlds. Algebra is the language of generality and structure, the art of manipulating symbols to uncover patterns and relationships. Vectors, on the other hand, provide the geometry of direction and magnitude—a way of describing the multi-dimensional spaces in which physical and conceptual systems evolve. Together, they offer a powerful framework for solving problems in mathematics, physics, engineering, and beyond. The origins of algebra stretch back to the ancient civilizations of Mesopotamia, Egypt, India, and Greece. Babylonian mathematicians (as early as 2000 BCE) developed methods to solve quadratic equations using geometric reasoning, though without symbolic notation. Over time, these ideas evolved into more abstract and generalized systems.

The true birth of algebra as a systematic discipline, however, is often credited to the Persian mathematician **Muḥammad ibn Mūsā al-Khwārizmī** (c. 780–850), whose name gave rise to the word *algorithm*. Working in the House of Wisdom in Baghdad during the Islamic Golden Age, Al-Khwarizmi wrote the seminal text *Al-Kitab al-Mukhtaṣar fi Hisab al-Jabr wal-Muqabala* ("The Compendious Book on Calculation by Completion and Balancing"), around the year 820. This treatise gave algebra its name (from *al-jabr*, meaning "reunion of broken parts") and presented systematic methods for solving linear and quadratic equations. His work emphasized procedures over symbolism, but it marked a turning point in mathematical thinking: problems could be abstracted, generalized, and solved by rule.

As algebra spread into medieval Europe via translations from Arabic into Latin, it evolved in form and sophistication. During the Renaissance, algebra began to shed its rhetorical character in favor of symbolic expression. **François Viète** (1540–1603) introduced a consistent use of letters to represent both known and unknown quantities, a crucial step toward modern notation. Later, **René Descartes** (1596–1650) merged algebra with geometry in his development of coordinate geometry, enabling equations to represent curves in space.

By the 19th century, algebra matured into a study not only of numbers and equations but of abstract structures. Mathematicians such as **Évariste Galois** (1811–1832) and **Niels Henrik Abel** (1802–1829) explored permutations, symmetries, and transformations—laying the groundwork for modern *abstract algebra*, including group, ring, and field theory. These frameworks now underpin much of modern mathematics, from cryptography to topology.

While algebra concerns itself with operations and relations, vectors are fundamentally geometric: they describe quantities that have both magnitude and direction. The intuition behind vectors can be traced to physical concepts such as velocity and force, but the formal mathematical treatment of vectors developed much later. In the 19th century, **Giusto Bellavitis** (1803–1880) introduced the notion of "equipollent

segments"—precursors to modern vectors. Around the same time, **William Rowan Hamilton** (1805–1865) formulated quaternions, a four-dimensional number system that extended complex numbers and described rotations in space. Though quaternions were later superseded by vector algebra in most practical contexts, they represented a major breakthrough in representing spatial transformations.

It was through the efforts of **Josiah Willard Gibbs** (1839–1903) and **Oliver Heaviside** (1850–1925) that vector analysis was formalized in a way accessible to engineers and physicists. Their introduction of the modern operations of vector addition, scalar multiplication, dot product, and cross product allowed vectors to become a standard language in electromagnetism, mechanics, and eventually computer graphics and machine learning. Today, algebra and vectors are not isolated tools but part of a unified mathematical language. Vectors inhabit *vector spaces*, which are studied using *linear algebra*, a field that combines algebraic rules with geometric insight. Concepts such as span, linear independence, dimension, and eigenvectors allow us to understand systems of equations, matrix transformations, and high-dimensional data with clarity and precision. Algebra provides structure; vectors give form. Together, they allow us to model change, symmetry, and interaction in both abstract and physical systems. From quantum states to economic models, from neural networks to mechanical forces, the union of algebraic reasoning and vector geometry continues to shape the mathematical description of the universe [...]. They remain as vital and foundational today as when their principles were first discovered, reflecting a timeless logic that underpins both human thought and the natural world.

A.2 Vectors and their properties

We will now review, very generally, the modern idea of vector, and its basic algebraic properties. Building upon the definitions of Gibbs and Heaviside, we will just extend the intuition of a vector as a list of numbers. According to the academic background, there are mainly three definitions of vector that are ofter found. From a physicist point of view, a vector is an arrow in space, with some information about length and direction. From a computer scientist perspective, a vector is just a list of elements. Let us define two vectors in $\mathbb{R}^3$ as column vectors:

$$\mathbf{v}_1 = \begin{pmatrix} v_{1x} \\ v_{1y} \\ v_{1z} \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} v_{2x} \\ v_{2y} \\ v_{2z} \end{pmatrix} \tag{A.1}$$

But from a mathematical, more general point of view, vectors are defined just as *element of a vector space*. This may sound abstract at first, but by *vector space* we just mean a set of elements that follow a given set of properties.

- Linearity: for any pair of vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, $f(\mathbf{v}_1 + \mathbf{v}_2) = f(\mathbf{v}_1) + f(\mathbf{v}_2)$
- Scaling: given a vector $\mathbf{v}$ and a scalar a, we can define the product, $a \cdot \mathbf{v} = \begin{pmatrix} a \cdot v_x \\ b \cdot v_y \\ c \cdot v_z \end{pmatrix}$

So, any element that satisfies these two rules, which is linear and can be multiplied by a scalar would be a vector [...]. Let's now see how to write down a vector in a proper way. Normally, we are very used to seeing vectors written in terms of their *components*, as follows.

$$\mathbf{v}_= \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}, \tag{A.2}$$

but here we are implicitly writing these components in reference to something. In particular, that description is writing $\mathbf{v}_1$ in terms of the *base*:

$$\mathbf{i} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{j} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{k} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{A.3}$$

Such that when we write $\mathbf{v}_1$ we are actually writing

$$\mathbf{v} = v_x \cdot \mathbf{i} + v_y \cdot \mathbf{j} + v_z \cdot \mathbf{k} = \mathbf{v}_x \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \mathbf{v}_y \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \mathbf{v}_z \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} v_{1x} \\ v_{1y} \\ v_{1z} \end{pmatrix} \tag{A.4}$$

Normally, we use the names x, y, z to denote the three cartesian axes, and i, j, k for the unitary vectors of the base. Un of the powerful features of writing vectors in this way is that we can now extend it to any number of dimensions, far beyond the 2D plane of the 3D space, and manipulate vectors in an arbitrarily large number of dimensions, n.

$$\mathbf{v} = v_1 \cdot e_1 + v_2 \cdot e_2 + ... + v_n \cdot e_n = \sum_{i=0}^{n} v_i \cdot e_i, \tag{A.5}$$

where v_i are the components, and e_i the vectors of the base. Note that the vectors of the base to satisfy two elementary properties. They are orthogonal to each other, and they they are unitary [...], and hence we call this an *orthonormal* base.

To compute the distance, or alignment between two vectors, we can define the dot product:
The dot product (also called the scalar product) of $\mathbf{v}_1$ and $\mathbf{v}_2$ is given by:

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = v_{1x}v_{2x} + v_{1y}v_{2y} + v_{1z}v_{2z} \tag{A.6}$$

This results in a scalar and measures the degree of alignment between the two vectors.

A similar quantity that describes the alignment is the cross product. The cross product (or vector product), defined only in three dimensions, yields a vector orthogonal to both $\mathbf{v}_1$ and $\mathbf{v}_2$:

$$\mathbf{v}_1 \times \mathbf{v}_2 = \begin{pmatrix} v_{1y}v_{2z} - v_{1z}v_{2y} \\ v_{1z}v_{2x} - v_{1x}v_{2z} \\ v_{1x}v_{2y} - v_{1y}v_{2x} \end{pmatrix} \tag{A.7}$$

This resulting vector is perpendicular to the plane defined by $\mathbf{v}_1$ and $\mathbf{v}_2$, with magnitude equal to the area of the parallelogram they span.

A.3 Matrices and linear transformations

Once we understand the idea of vector, base and linear combination, we can try to describe movements, rotations and transformations, just by thinking how do they affect the vectors of the base. This is why transformations, or *operators*, are represented as matrices.

$$\mathcal{O} \cdot \mathbf{v}_1 = \begin{pmatrix} O_{11}O_{12}O_{13} \\ O_{21}O_{22}O_{23} \\ O_{31}O_{32}O_{33} \end{pmatrix} \cdot \begin{pmatrix} \cdot v_x \\ \cdot v_y \\ \cdot v_z \end{pmatrix} = \begin{pmatrix} O_{11}v_x + O_{12}v_y + O_{13}v_z \\ O_{21}v_x + O_{22}v_y + O_{23}v_z \\ O_{31}v_x + O_{32}v_y + O_{33}v_z \end{pmatrix} \tag{A.8}$$

A.4 Basic algebraic operations

Basic algebraic operations

Appendix B

Appendix: Functions and derivatives: a quick review

B.1 From curves to calculus: functions

At its heart, calculus is the mathematics of change. It arose not from abstraction alone but from an urgent need to describe motion, growth, and the ever-shifting patterns of the natural world. Long before the formalism of equations and limits, ancient civilizations—including the Greeks, Indians, and Chinese—grappled with questions of geometry, area, and continuity. The Greek philosopher Zeno of Elea (5th century BCE) famously posed paradoxes about motion and infinity—such as the arrow that never reaches its target or Achilles who cannot overtake the tortoise—highlighting the conceptual difficulties of change and division.

The seeds of calculus began to germinate more concretely in the late medieval and early Renaissance periods. The Indian mathematician Bhāskara II (12th century) hinted at the idea of instantaneous rates of change, and later, the French philosopher René Descartes (1596–1650) and the English mathematician John Wallis (1616–1703) laid important groundwork in algebra and infinite series. But calculus, as a coherent system, was born in the late 17th century through the independent and nearly simultaneous efforts of **Isaac Newton** (1643–1727) in England and **Gottfried Wilhelm Leibniz** (1646–1716) in Germany. Newton, motivated by problems in physics—such as planetary motion and gravitation—developed what he called "the method of fluxions," focusing on rates of change and motion. Leibniz, meanwhile, introduced a notational system and conceptual clarity that would become the foundation for the modern symbolic language of calculus. His elegant integral sign $\int$ and differential notation dx, dy remain standard to this day.

Central to the language of calculus is the concept of a **function**—a rule or relationship that assigns to each input a single output. While the basic idea can be traced back to ancient times (e.g., geometric ratios in Euclid), the function began to take formal shape in the 17th and 18th centuries. Descartes' invention of analytic geometry (linking algebra to geometry) made it possible to graph equations as curves. Later, **Leonhard Euler** (1707–1783), perhaps the most prolific mathematician of all time, introduced the modern notation for functions, such as $f(x)$, and explored their properties across a wide range of mathematical and physical contexts. Functions became the foundation for modeling everything from the arc of a cannonball to the vibrations of a violin string.

In calculus, we study how functions behave, how they change, and how we can understand their subtleties through two great operations: **differentiation** and **integration**. Differentiation captures the idea of instantaneous change—the speed of a falling apple or the slope of a curve at a single point. Newton used it to describe acceleration and the motion of celestial bodies, revolutionizing classical mechanics. Integration, on the other hand, concerns accumulation—how quantities build up over time or space. The method of exhaustion used by **Archimedes** (287–212 BCE) was an early form of integration, but it was Newton and Leibniz who unified the concept, showing that differentiation and integration are inverse processes—a

profound insight now known as the **Fundamental Theorem of Calculus**.

At its heart, calculus is the mathematics of change. It arose not from abstraction alone but from an urgent need to describe motion, growth, and the ever-shifting patterns of the natural world. Long before the formalism of equations and limits, ancient civilizations—including the Greeks, Indians, and Chinese—grappled with questions of geometry, area, and continuity. The Greek philosopher Zeno of Elea (5th century BCE) famously posed paradoxes about motion and infinity—such as the arrow that never reaches its target or Achilles who cannot overtake the tortoise—highlighting the conceptual difficulties of change and division.

In calculus, we study how functions behave, how they change, and how we can understand their subtleties through two great operations: **differentiation** and **integration**. Differentiation captures the idea of instantaneous change—the speed of a falling apple or the slope of a curve at a single point. Newton used it to describe acceleration and the motion of celestial bodies, revolutionizing classical mechanics. Integration, on the other hand, concerns accumulation—how quantities build up over time or space. The method of exhaustion used by **Archimedes** (287–212 BCE) was an early form of integration, but it was Newton and Leibniz who unified the concept, showing that differentiation and integration are inverse processes—a profound insight now known as the **Fundamental Theorem of Calculus**.

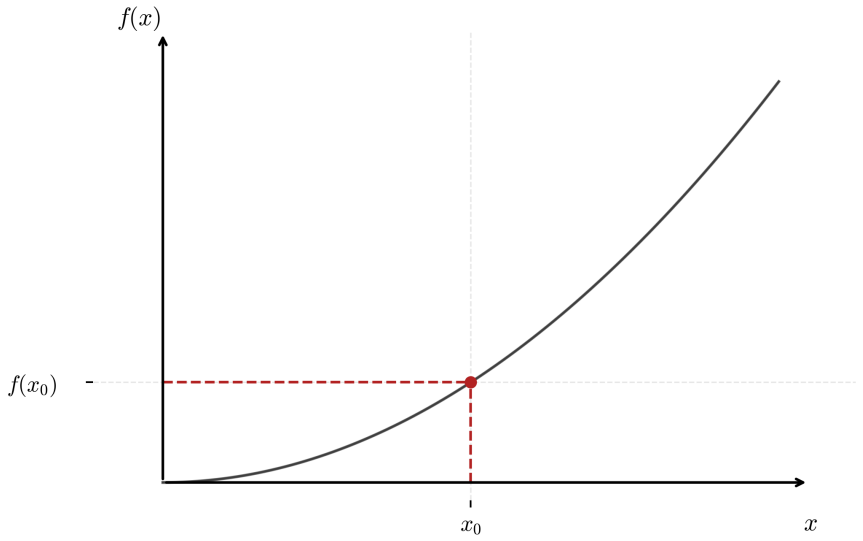


Figure B.1: Representation of a function $f(x)$, a given point x_0 and its image $f(x_0)$.

The seeds of calculus began to germinate more concretely in the late medieval and early Renaissance periods. The Indian mathematician Bhāskara II (12th century) hinted at the idea of instantaneous rates of change, and later, the French philosopher René Descartes (1596–1650) and the English mathematician John Wallis (1616–1703) laid important groundwork in algebra and infinite series. But calculus, as a coherent system, was born in the late 17th century through the independent and nearly simultaneous efforts of **Isaac Newton** (1643–1727) in England and **Gottfried Wilhelm Leibniz** (1646–1716) in Germany. Newton, motivated by problems in physics—such as planetary motion and gravitation—developed what he called "the method of fluxions," focusing on rates of change and motion. Leibniz, meanwhile, introduced a notational system and conceptual clarity that would become the foundation for the modern symbolic language of calculus. His elegant integral sign $\int$ and differential notation dx, dy remain standard to this day.

B.2 Change, slope and minima: derivatives

Once we are familiar with the idea of functions as a way of representing the relation between two variables, we can try to model how functions *change*. As already mentioned, this builds upon the work of Newton and Leibniz, although the actual mathematical language we will use was developed much later. Consider we have a function $f(x)$, like the one represented in the figure below. Regardless of the specific shape, or dependency, of $f(x)$ we could pick a point x_0, and move a small amount Δx along the x axis, to $x_0 + \Delta x$. At this point, we could compute how much that change impacts the y axis, just by evaluating $f(x_0)$, and after the displacement $f(x_0 + \Delta x)$

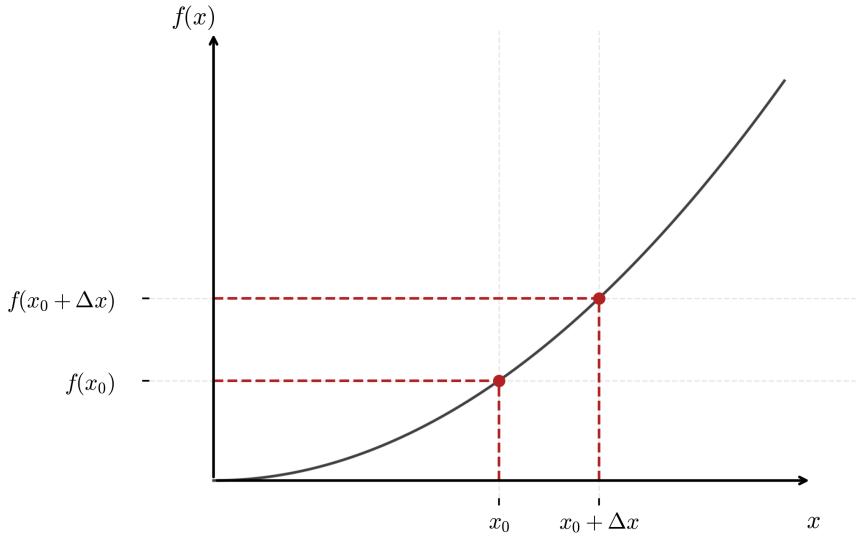


Figure B.2: Representation of a function $f(x)$, a given point x_0 and its image $f(x_0)$, and an increment Δx from x_0 to $x_0 + \Delta x$. The ratio between the increments in the horizontal axis, Δx, and the vertical axis $f(x_0 + \Delta x) - f(x_0)$ represents the derivative of the function at that point, that we denote by $f'(x_0)$.

To evaluate how fast a function grows - or decreases, to capturing the idea of *change*, we could just compute the ratio if the difference Δx, and the corresponding change in $f(x)$,

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} \quad . \tag{B.1}$$

The derivative of f at a point x_0 is defined just as that ratio, for an infinitesimally small Δx:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{B.2}$$

This expression represents the slope of the tangent line to $f(x)$ at the point x_0. If the limit exists, f is said to be *differentiable* at x. It measures how a function responds to small variations in its input. Intuitively, if a function $f(x)$ describes a quantity, then its derivative $f'(x)$ tells us how fast that quantity is changing at each point. There are several notations for the derivative, listed as follows:

$$f'(x), \quad \frac{df}{dx}, \quad \frac{dy}{dx} \quad (\text{if } y = f(x))$$

Each emphasizes a slightly different aspect. The first notation, attributed to Newton, represents $f'(x)$ as a function, while Leibniz notation df/dx, express it as a rate of change, the ratio of increments along the function Δf over Δx. Notations may change across books and literature sources; here we will denote Δ an increment of arbitrary size, and we will call it *infinitesimal*, or *differential*, when we write it as dx

Hence, we can characterize the equilibrium - or *stationary* - points of a function, its maxima and minima, just by the condition

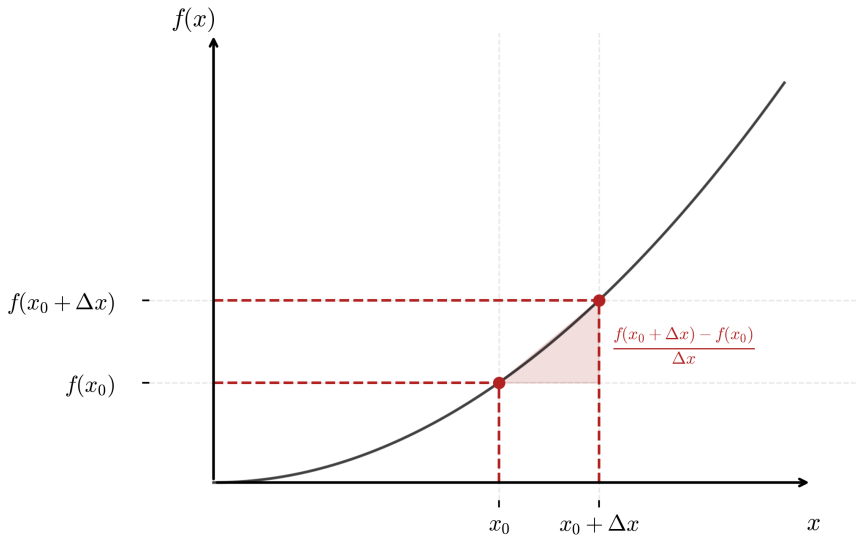


Figure B.3: Representation of a function $f(x)$, a given point x_0 and its image $f(x_0)$, and an increment Δx from x_0 to $x_0 + \Delta x$. The ratio between the increments in the horizontal axis, Δx, and the vertical axis $f(x_0 + \Delta x) - f(x_0)$ represents the derivative of the function at that point, that we denote by $f'(x_0)$.

$$f'(x) = 0 \quad . \tag{B.3}$$

Wherever a change along x does not affect $f(x)$, meaning $f(x_0 + \Delta x) - f(x_0) = 0$, will be either a maximum or a minimum, and we know that there is no growth around that point x_0.

The development of calculus marked a turning point in intellectual history. It gave scientists, engineers, and economists the tools to model the world with precision and predict its behavior with confidence. The mathematical machinery behind electricity, fluid dynamics, statistics, and quantum mechanics all rely on the principles laid down by Newton, Leibniz, Euler, and their successors. Yet beyond its utility, calculus offers something more enduring: a glimpse into the deep continuity that underlies the universe, where motion and form, quantity and change, are all expressions of a single, evolving logic. Through calculus and the language of functions, we do not merely calculate—we begin to understand.

Geometrically, the derivative represents the slope of the tangent line to the curve $y = f(x)$. Physically, if $f(t)$ describes the position of a moving object as a function of time, then $f'(t)$ gives its instantaneous velocity.

Example

If $f(x) = x^2$, then:

$$f'(x) = \lim_{\Delta x \to 0} \frac{(x + \Delta x)^2 - x^2}{\Delta x} = \lim_{\Delta x \to 0} \frac{2x\Delta x + (\Delta x)^2}{\Delta x} = 2x$$

Thus, the derivative of x^2 is $2x$, meaning the rate of change increases linearly with x.

The development of calculus marked a turning point in intellectual history. It gave scientists, engineers, and economists the tools to model the world with precision and predict its behavior with confidence. The mathematical machinery behind electricity, fluid dynamics, statistics, and quantum mechanics all rely on the principles laid down by Newton, Leibniz, Euler, and their successors. Yet beyond its utility, calculus offers something more enduring: a glimpse into the deep continuity that underlies the universe, where motion and form, quantity and change, are all expressions of a single, evolving logic. Through calculus and the language

of functions, we do not merely calculate—we begin to understand.

B.3 Divergence and gradient: differential calculus

Divergence and gradient: differential calculus

Appendix C

Appendix: Integrals: a quick review

Integral calculus is the mathematics of accumulation and area, of summing infinitesimal parts to understand wholes. Where differential calculus dissects change, integral calculus assembles continuity—weaving together countless small contributions to reveal the total. It addresses fundamental questions: How much space does a shape occupy? How much work is done over a path? How do quantities build up over time? From measuring land in antiquity to solving complex equations in physics, integral calculus has served as a bridge between the discrete and the continuous, the finite and the infinite.

The Historical Origins of Integration. The intuitive idea of integration—measuring area under a curve or volume under a surface—dates back to the ancient world. The Greek mathematician **Eudoxus of Cnidus** (c. 390–337 BCE) developed the *method of exhaustion*, a precursor to integration, which used successively refined polygons to approximate areas and volumes. **Archimedes** (c. 287–212 BCE), perhaps the greatest of ancient mathematicians, applied this method to derive results such as the area of a parabola segment and the volume of a sphere, anticipating integral calculus by nearly two millennia

Yet it was not until the 17th century that integration found its modern form. The crucial turning point came with the realization that integration and differentiation are inverse processes. This insight was crystallized by **Isaac Newton** (1642–1727) and **Gottfried Wilhelm Leibniz** (1646–1716), working independently. Newton, motivated by problems in physics, viewed integration as the accumulation of continuous quantities like motion and force. Leibniz, meanwhile, formalized the symbolic language of integration, introducing the elongated S symbol ($\int$) for summation and dx for infinitesimal change. His notational clarity profoundly shaped the development of analysis.

Their work established what we now call the **Fundamental Theorem of Calculus**, which states that integration and differentiation are inverse operations:

$$\frac{d}{dx}\left(\int_a^x f(t)\,dt\right) = f(x) \qquad \text{and} \qquad \int_a^b f'(x)\,dx = f(b) - f(a)$$

This theorem provided both the conceptual unity and practical tools to solve problems in geometry, physics, and beyond.

Rigor and Expansion. Though powerful, early calculus lacked rigor. Questions arose: What precisely is a limit? What functions can be integrated? These were resolved over the 18th and 19th centuries by mathematicians such as **Augustin-Louis Cauchy** (1789–1857), who formalized limits and continuity, and **Bernhard Riemann** (1826–1866), who defined integration through Riemann sums, grounding the intuitive notion of area in precise terms. Later, **Henri Lebesgue** (1875–1941) expanded the theory further with the *Lebesgue integral*, which could handle more pathological functions and laid the foundation for modern measure theory. This broader framework enabled integration in more abstract contexts, from probability spaces to Hilbert spaces, and opened the door to modern analysis and functional integration.

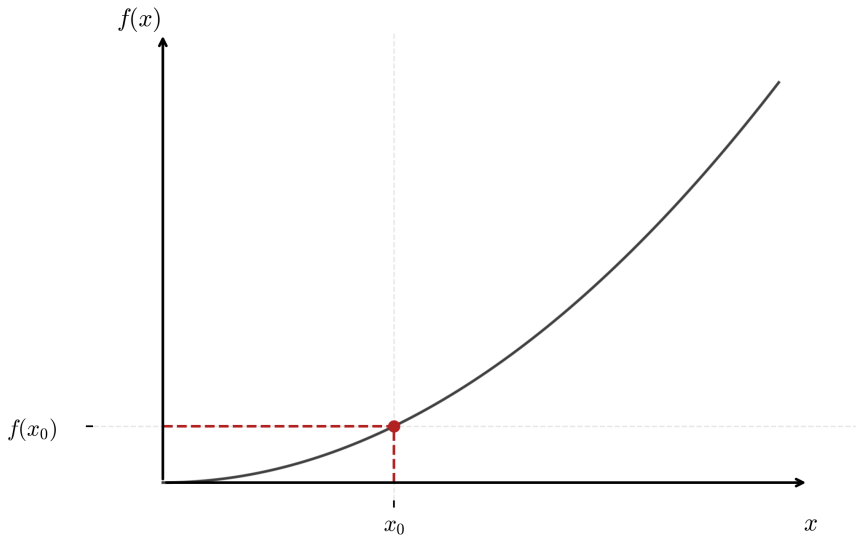


Figure C.1: Representation of a function $f(x)$, a given point x_0 and its image $f(x_0)$, and an increment Δx from x_0 to $x_0 + \Delta x$.

Together, these statements tell us that the process of summing infinitesimal contributions to find total accumulation (integration) and the process of finding the instantaneous rate of change (differentiation) are fundamentally linked. This profound insight lies at the heart of calculus, enabling the solution of countless problems in physics, engineering, and beyond.

Applications and Legacy.Integral calculus plays a central role in virtually every branch of science and engineering. In physics, it describes the flow of fluids, the propagation of waves, the accumulation of charge, and the work done by forces over time. In economics, it models total cost, revenue, and utility. In biology and medicine, it helps quantify rates of growth, spread of populations, and diffusion processes.

Computationally, integration techniques underpin numerical methods, algorithms for simulations, and digital signal processing. In probability theory, integrals define expected values and distributions, while in geometry and topology, integration is used to calculate lengths, areas, and volumes in curved and abstract spaces.

The Spirit of Integral Calculus. Integral calculus teaches us to think holistically—to understand how global behavior arises from local contributions. It invites us to see curves not just as paths but as stories of accumulation: of distance traveled, energy stored, substance gathered. From the earliest geometric approximations to the abstract theories of modern analysis, integration has grown into one of the most powerful and beautiful tools in mathematics.

To learn integral calculus is to acquire a lens through which we view the continuous world—to perceive not only rates of change, but the subtle ways in which all things are built up from infinitesimal parts. It is a mathematics of synthesis, binding motion to rest, parts to whole, and analysis to intuition.

C.1 Indefinite integral as antiderivative

Integral calculus complements the derivative by focusing on *accumulation* rather than instantaneous change. It allows us to find areas under curves, total quantities accumulated over an interval, and much more.

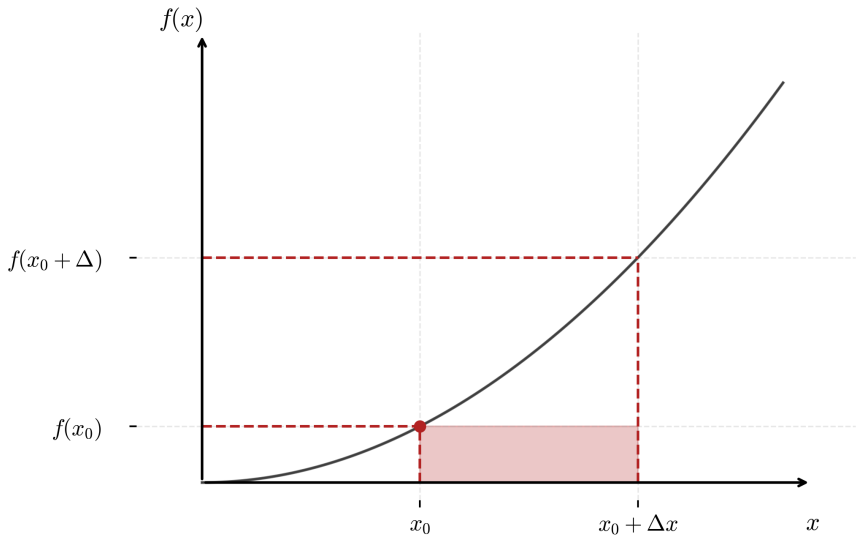


Figure C.2: Representation of a function $f(x)$, a given point x_0 and its image $f(x_0)$, and an increment Δx from x_0 to $x_0 + \Delta x$. Here we divide the Δx increment in smaller steps, aiming to approximate the area under the function. As the subdivisions become smaller, they become a better approximation of the area, and in the limit $\Delta x \to \infty$ they converge to the Riemann definition.

Definite Integral

Given a function $f(x)$ defined on an interval $[a, b]$, the definite integral of f from a to b is intuitively the total accumulation of the values of f over that interval. It is formally defined as the limit of Riemann sums:

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\,\Delta x_i$$

where the interval $[a, b]$ is partitioned into subintervals of length Δx_i, and x_i^* is a sample point in the i-th subinterval. This sum approximates the area under the curve $y = f(x)$ between $x = a$ and $x = b$.

The Fundamental Theorem of Calculus bridges differentiation and integration, showing they are inverse processes. It has two parts:

- **Part 1:** If $F(x)$ is defined by
$$F(x) = \int_a^x f(t)\,dt,$$
where f is continuous on $[a, b]$, then F is differentiable and
$$F'(x) = f(x).$$
In other words, differentiation "undoes" integration.

- **Part 2:** If F is any antiderivative of f on $[a, b]$ (i.e., $F'(x) = f(x)$), then
$$\int_a^b f(x)\,dx = F(b) - F(a).$$
This means the definite integral can be evaluated by finding any antiderivative of f.

C.2 Define definite integral as area under a curve

Define definite integral as area under a curve

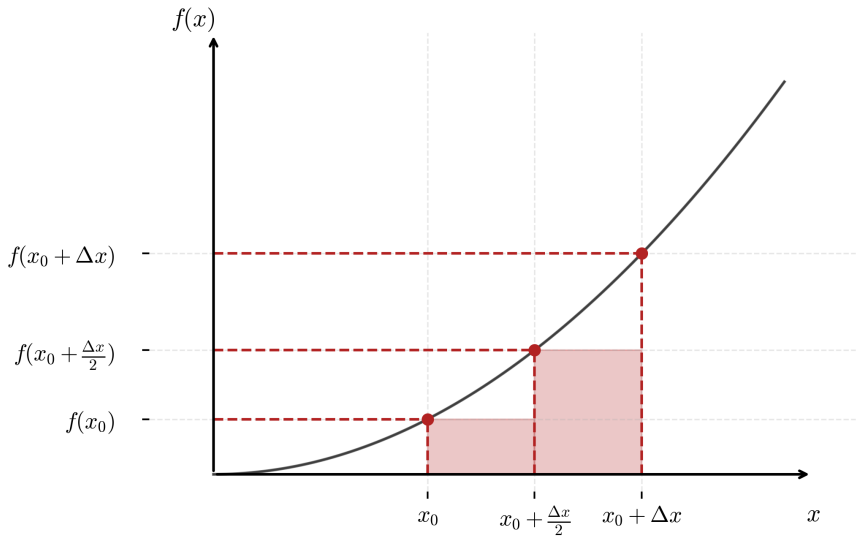


Figure C.3: Representation of a function $f(x)$, a given point x_0 and its image $f(x_0)$, and an increment Δx from x_0 to $x_0 + \Delta x$. Here we divide the Δx increment in smaller steps, aiming to approximate the area under the function. As the subdivisions become smaller, they become a better approximation of the area, and in the limit $\Delta x \to \infty$ they converge to the Riemann definition.

C.3 The fundamental theorem of calculus

The fundamental theorem of calculus

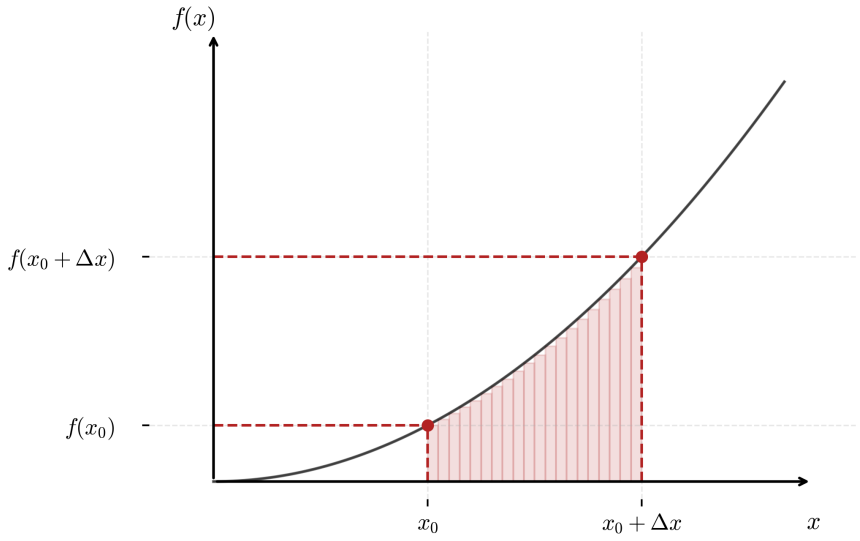


Figure C.4: Representation of a function $f(x)$, a given point x_0 and its image $f(x_0)$, and an increment Δx from x_0 to $x_0 + \Delta x$. Here we divide the Δx increment in smaller steps, aiming to approximate the area under the function. As the subdivisions become smaller, they become a better approximation of the area, and in the limit $\Delta x \to \infty$ they converge to the Riemann definition.

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* (4th ed.). Pearson, 2012.

[4] P.S. Bandyopadhyay and M.R. Forster (Eds.). *Philosophy of Statistics*. Handbook of the Philosophy of Science, Vol. 7, Elsevier, 2011.

[5] M. Diez, D. Barr, and 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. *Ancient Board Games in Perspective*. British Museum Press, 2007.

[8] Florence N. David. *Games, Gods and Gambling: A History of Probability and Statistical Ideas*. Charles Griffin & Company, 1962.

[9] Marcus Tullius Cicero. *De Natura Deorum & De Divinatione*. Loeb Classical Library (various editions), ca. 45 BCE.

[10] Gerolamo Cardano. *Liber de Ludo Aleae (Book on Games of Chance)*. Posthumously published 1663; English translation in Oystein Ore, *Cardano, The Gambling Scholar*, Princeton University Press, 1953.

[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*. Published in Latin, 1657.

[13] Jacob Bernoulli. *Ars Conjectandi*. Thurnisius, 1713. English translation by Edith Dudley Sylla, Johns Hopkins University Press, 2006.

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

[15] Andrey N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933. English translation: *Foundations of the Theory of Probability*, Chelsea Publishing Company, 1956.

[16] Ramsey, F. P. (1926). *Truth and Probability*. In D. H. Mellor (Ed.), *The Foundations of Mathematics and Other Logical Essays* (pp. 156–198). London: Routledge & Kegan Paul.

[17] Pearson, K. (1892). *The Grammar of Science*. London: Adam and Charles Black. (Contains the introduction of the term and concept of histogram).

[18] Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley. (First book introducing the box plot as a graphical method).

[19] Hintze, J.L., & Nelson, R.D. (1997), *Violin Plots: A Box Plot-Density Trace Synergism*. The American Statistician, 52(2), 181–184. doi:10.1080/00031305.1998.10480559 (Introduced the violin plot in statistics).

[20] Anscombe, F. J. (1973), *Graphs in statistical analysis*. The American Statistician, 27(1), 17–21.

[21] Spearman, C. (1904), *The proof and measurement of association between two things*. The American Journal of Psychology, 15(1), 72–101.

[22] T. Bayes, "An Essay towards solving a Problem in the Doctrine of Chances," (1763).

[23] P. Billingsley, *Probability and Measure* (1995).

[24] B. de Finetti, *Theory of Probability* (1974).

[25] J. L. Doob, *Stochastic Processes* (1953).

[26] A. N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung* (1933).

[27] P.-S. Laplace, *Théorie analytique des probabilités* (1812).

[28] H. Lebesgue, *Intégrale, longueur, aire* (1902).

[29] R. von Mises, *Wahrscheinlichkeit, Statistik und Wahrheit* (1928).

[30] C. Huygens, *De ratiociniis in ludo aleae* (1657).

[31] J. Bernoulli, *Ars Conjectandi* (1713).

[32] P.-S. Laplace, *Théorie analytique des probabilités* (1812).

[33] S.-D. Poisson, *Recherches sur la probabilité des jugements* (1837).

[34] Gerolamo Cardano. *Liber de Ludo Aleae (Book on Games of Chance)*. Posthumously published 1663; English translation in Oystein Ore, *Cardano, The Gambling Scholar*, Princeton University Press, 1953.

[35] C. F. Gauss, *Theoria motus corporum coelestium* (1809).

[36] A. N. Kolmogorov, *Foundations of the Theory of Probability* (1933).

[37] K. Pearson, *Contributions to the Mathematical Theory of Evolution* (1895).

[38] H. Lebesgue, *Intégrale, longueur, aire*. Annali di Matematica, 1902.

[39] J. L. Doob, *Stochastic Processes*. Wiley, 1953.

[40] R. von Mises, *Probability, Statistics and Truth*, 1928.

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

[42] P.-S. Laplace, *Théorie Analytique des Probabilités*, 1812.

[43] B. de Finetti, *Theory of Probability*. Wiley, 1974.

[44] P. Billingsley, *Probability and Measure*. Wiley, 1995.

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

[46] Student. (1908). *The probable error of a mean*. Biometrika, 6(1), 1–25.

[47] Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

[48] Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

[49] Pearson, K. (1900). *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.* Philosophical Magazine Series 5, 50(302), 157–175.

[50] Welch, B. L. (1947), *The generalization of "Student's" problem when several different population variances are involved.* Biometrika, 34(1–2), 28–35.

[51] Yates, F. (1934), *Contingency tables involving small numbers and the χ^2 test.* Supplement to the Journal of the Royal Statistical Society, 1(2), 217–235.

[52] Greenwood, M., & Yule, G. U. (1911), *An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents.* Journal of the Royal Statistical Society, 83(2), 255–279.

[53] B. Pascal and P. de Fermat, *Correspondence on the theory of probability*, 1654.

[54] Ian Hacking, *The Logic of Statistical Inference*, Cambridge University Press, 1965.

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

[56] Pierre-Simon Laplace, *Essai philosophique sur les probabilités*, Paris: Courcier, 1814.

[57] William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, Wiley, 1950.

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

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