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

Jesús Urtasun Elizari

January 17, 2026

Contents

Preface iii
The purpose of these notes . . . iii

Introduction v
A bit of history . . . v

1 Descriptive statistics 1
1.1 Sampling and data types . . . 2
1.2 Central tendency and variation . . . 4
1.3 Data visualization . . . 7

2 Foundations of Probability 11
2.1 Probability and random events . . . 12
2.2 Discrete events . . . 13
2.2.1 Bernoulli trials . . . 13
2.2.2 Binomial distribution . . . 14
2.2.3 Poisson distribution . . . 14
2.2.4 Discrete uniform distribution . . . 15
2.3 Continuous events . . . 16
2.3.1 Gaussian distribution . . . 17
2.3.2 Exponential distribution . . . 17
2.3.3 Continuous uniform distribution . . . 18
2.4 Expected values . . . 19
2.5 Mean and variance of common distributions . . . 20

3 Estimation, variability and confidence 23
3.1 The Law of Large Numbers . . . 23
3.2 The Central Limit Theorem . . . 24
3.3 Bias, variance and Mean Squared Error . . . 24
3.4 Confidence intervals and critical regions . . . 24

4 Introduction to hypothesis testing 27
4.1 Prediction vs inference revisted . . . 28
4.2 General approach to hypothesis testing . . . 29
4.3 Statistic tests: common examples . . . 30
4.3.1 One sample t-test: Compare sample mean with hypothesized value . . . 30
4.3.2 Two sample t-test: Compare sample means of two independent groups . . . 32
4.3.3 Fisher's F test: Compare variation of two independent groups . . . 33
4.3.4 Fisher's ANOVA: Compare variation of multiple groups . . . 34
4.3.5 Pearson's χ^2 test: Compare distributions and testing for normality . . . 35
4.3.6 The Wald test: asymptotic behavior . . . 36
4.4 Parametric and non-parametric tests . . . 36

4.4.1 Wilcoxon signed-rank test 37
4.4.2 Mann–Whitney U test 37
4.4.3 Levene median-based test 37
4.4.4 Kruskal–Wallis test 37
4.4.5 The Kolmogorov–Smirnov test 38
4.4.6 The Shapiro–Wilk test 38
4.5 Error types in hypothesis testing 38

5 Introduction to conditional probability 44
5.1 Motivation and philosophy 44
5.2 Dependent and independent events 44
5.3 Some examples of conditional probability 44

A Appendix 1 46

B Appendix 2 47

C Appendix 3 48

Preface

The purpose of these notes

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

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

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

This is intended to be a complete introductory course, and no previous mathematical background is required. By keeping the theory simple and always followed by examples, we will build the definitions and quantities from simple to more complex. All mathematical formulas will be introduced with rigorous notation, but keeping in mind that it is not the symbols or the numbers, but the intuitions and the general understanding, what we are after. Additionally, all topics will be introduced alongside with some short historical discussion and context, as we believe that a purely technical knowledge just grasps the complexity—and beauty—of scientific topics. As one could anticipate already, a proper understanding of ideas such as uncertainty, variation, chance, probability, inference, etc, can be applied to describing a vast amount of real-world phenomena, ranging from gambling and to games of change to data analysis and modelling in physics, biology, machine learning and quantum mechanics, among many others.

As mentioned, the course is organised in five chapters.
Chapter 1 [...] Chapter 2 [...] Chapter 3 [...] Chapter 4 [...] Chapter 5 [...]

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

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

- A simple, intuitive introduction to statistics with few mathematical concepts is provided in Spiegelhalter's *"The Art of Statistics: How to Learn from Data"* [2].
- A more foundational textbook, with more advanced mathematical approach, can be found at DeGroot and Schervish's *"Probability and Statistics"* [3].
- For a philosophical and historical perspective on probability and statistics, please find Forster and Bandyopadhyay's handbook *"Philosophy of Statistics"* [4].
- A comprehensive introduction with focus on practical applications and modern data analysis tools is can be found at Diez, Barr & Mine *"OpenIntro Statistics"* [5].
- For fundamental concepts in probability and statistics, including random variables, distributions and statistical inference, with practical examples and exercises follow Hossein Pishro-Nik's *"Probability, Statistics & Random Processes"* [6].

Introduction

Even fire obeys the laws of numbers.
— J.B. Joseph Fourier

A bit of history

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

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

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

The formalization of probability as a mathematical discipline did not occur until the 17th century, most notably through the seminal correspondence between Blaise Pascal and Pierre de Fermat. Their work, motivated by problems such as finding a fair division of stakes in interrupted games of chance, introduced foundational concepts such as combinatorics, expected values, and variance [11]. These developments paved the way for later contributions by Christiaan Huygens, who in 1657 wrote the first published textbook on probability *"De Ratiociniis in Ludo Aleae"* [12], and Jacob Bernoulli, whose 1713 *"Ars Conjectandi"* remains among the most influential early texts in the field. Their works, along with many others, collectively laid the groundwork for the probabilistic and statistical methods that foreshadow modern scientific reasoning [13, 14].

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

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

The modern axiomatic formulation of probability was introduced in the early 20th century by the Russian mathematician Andrey Kolmogorov. In his 1933 monograph *"Grundbegriffe der Wahrscheinlichkeitsrechnung"* (*"Foundations of the Theory of Probability"*) [15], Kolmogorov synthesized classical and frequentist ideas into a rigorous mathematical framework based on measure theory. His axioms remain the standard foundation for probability theory to this day. It may seem surprising that a concept with such ancient origins was not formally axiomatized until relatively recent times, and we will return to Kolmogorov's formulation and its implications in greater detail in Chapter 5. Nevertheless, philosophical discussions about the interpretation of probability and its relation to the physical sciences—especially in the context of determinism, epistemology, and modern topics such as quantum mechanics—predate Kolmogorov's formulation and continue to evolve to this day.

Chapter 1

Descriptive statistics

Statistics is the grammar of science.

— Karl Pearson

As a first approach to probabilty and statistics, we should properly define both topics and their main fields of study. Even deeply related, and both rooted in *combinatorics*—the study of uncertainty and things that change—they constitute well differentiated fields of mathematical analysis. A clear distinction often made is that probability is a *predictive* branch of mathematics, dealing with random events, also referred to as *stochastic*, aiming to compute expected values for such unknown outcomes. On the other hand, statistics would be a *descriptive* way of dealing with uncertainty, by sampling finite sets of observations from a given population, and building informative quantities, called statistical *estimators* to explore central tendency and variation. Such distinction has been extensively debated and discussed by mathematicians, experimental scientists, and philosoplers of science.

As a rule of thumb, probability provides a formal language for modelling uncertainty, whereas statistics concerns the epistemic problem of learning from data. Through this chapter we will introduce basic ideas on statistical inference such as population, sampling, and estimators of central tendency and variation, together with some notions of representation and visualization. The foundations of probability theory, rooted in the works of Bernoulli, Laplace, and Gauss, among others, will be covered in Chapter 2. Hence, a philosophical position often adopted is that statistics is essentially the study of uncertainty, and that the statistician's role is to assist other fields who encounter uncertainty in their work. In practice, there is a restriction in that statistics is ordinarily associated with data; and it is the link between the uncertainty, or variability, in the data and that in the topic itself that has occupied statisticians. Statistics does not have a monopoly of studies of uncertainty. Probability discusses how randomness in one part of a system affects other parts.

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

1.1 Sampling and data types

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

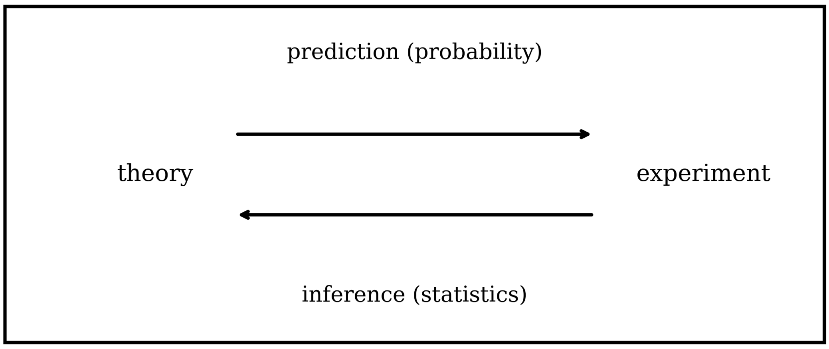


Figure 1.1: Representation of the predictive (from theory, or model, to experimental verification) and inferential (from data, measurement, observation to underlying truth) approaches to natural phenomena. As an example of the predictive branch of mathematics dealing with uncertainty we would find the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

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

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

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

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

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

where the elements the sample x_i consist of just a selected group of observations from the population, not necessarily consecutive or in the same order. The population represents the ideal object of inference, while the sample is the concrete, finite evidence available to us. As an example, if I want to study some disease and its relation smokers in a given country, I will never have access to the *complete population*, but only the amout of them that I am able to question, measure, or survey. This distinction is far from trivial. A poorly chosen sample often misrepresents the population and may induce bias, whereas a carefully constructed one mirrors its essential features, and can be used to describe the underlying nature.

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

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

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

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

Let's end this section with a historical note. As we have mentioned, uncertainty has been associated with games of chance and gambling from quite old times, but it was not adressed as a statistics problem until much later. The Royal Statistical Society, founded in 1834, together with many other statistical groups, was originally set up to just gather and publish data, as an attempt to reduce such uncertainty. It did not take long before statisticians wondered how the data might best be used and modern *statistical inference* was born. Charles Babbage, Adolphe Quetelet [...]. Among its famous members was Florence Nightingale, the society's first female member in 1858, whose work was shaped by this same intellectual climate. [...] Other notable RSS presidents have included William Beveridge, Ronald Fisher, which we will discuss in Chapter 4.

Andrew Lang's famous quote *"most people use statistics as a drunken man uses lamp-posts—for support rather than illumination"*, highlights the tendency to use statistics as a crutch, relying on them for validation rather than seeking genuine understanding. Lang's observation serves as a cautionary reminder to approach statistical data with critical thinking and not merely as a tool to bolster preconceived notions.

1.2 Central tendency and variation

Once we have a clear distinction between the population under study and the selected sample, we face a problem. Neither the population—referred as the *true*—mean value, sometimes written as μ, nor its variance–referred as the *true* variance, and written as σ^2 are available to us. As we just saw, the *only thing we have is the finite set of observations in our sample*, hence we could try to build some "informative quantities" out of out data that would give us a hint of the central value, a measure of spread, etc. Such quantities are called *statistical estimators.* Common examples of such estimators are the *sample mean*, the *median*, and the *variance*, among others.

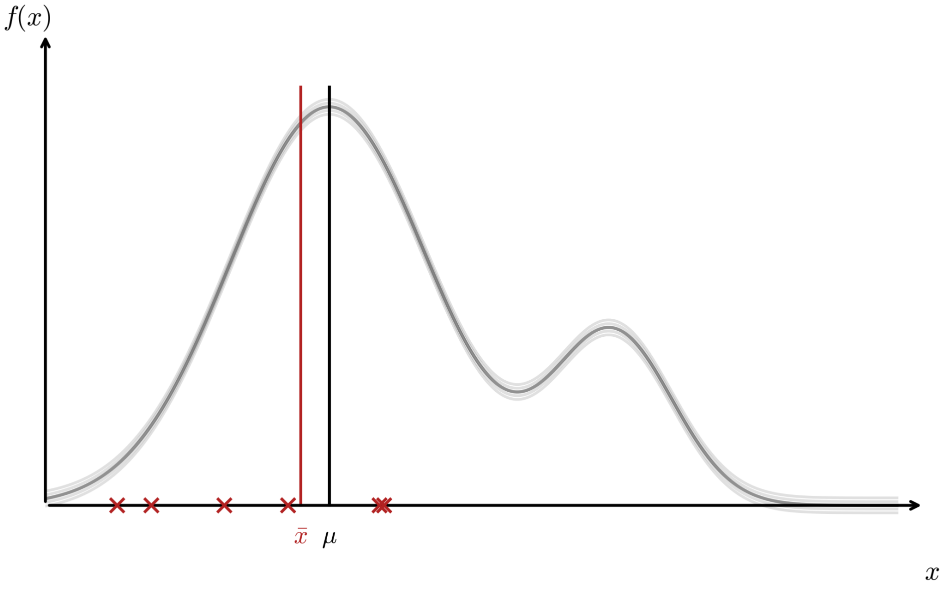


Figure 1.2: Representation of the *true* population mean μ, in black, and the observed *sample* mean $\bar{x}$. The true mean is and ideal and unaccessible quantity, while the sample mean can be computed as an estimator of the finite sample.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$M = 3 \ .$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Finally, *quantiles* divide the ordered data into equal proportions. The p-th quantile Q_p is the value below which a fraction p of the data lies. Special cases include the *first quartile* (Q_1, 25th percentile), the *median* (Q_2, 50th percentile), and the *third quartile* (Q_3, 75th percentile). A rigurous definition of quantiles requires the idea of distribution and cumulative probability, so we will discuss them in next chapter. As a note, for a continuous cumulative distribution function (CDF) F, the p-th quantile satisfies

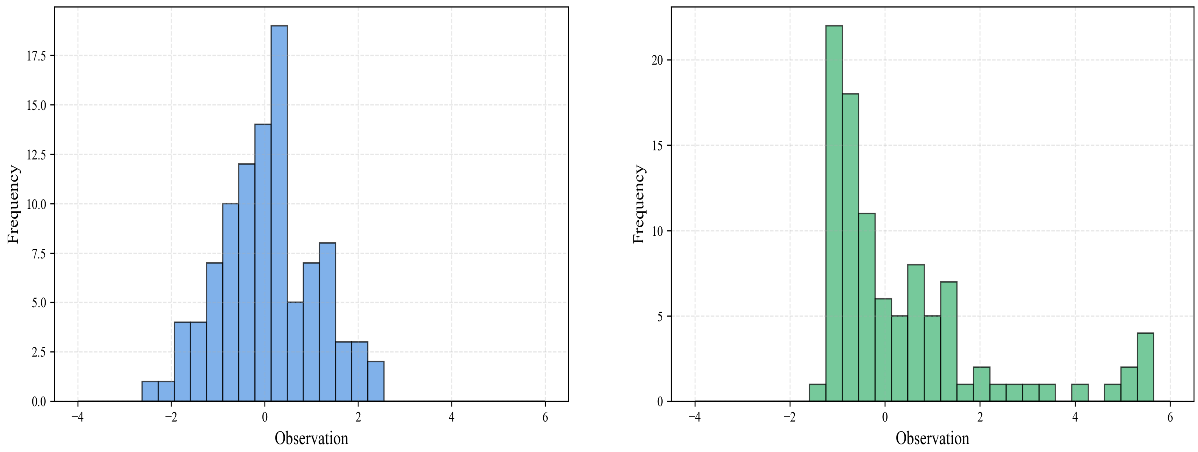
$$Q_p = \inf\{x : F(x) \geq p\}. \tag{1.8}$$

In summary, mean, median, mode, variance, standard deviation, and quantiles provide a rich, complementary view of the dataset's central tendency and variability, allowing for both numerical and graphical summaries that capture the essence of the data.

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

1.3 Data visualization

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



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

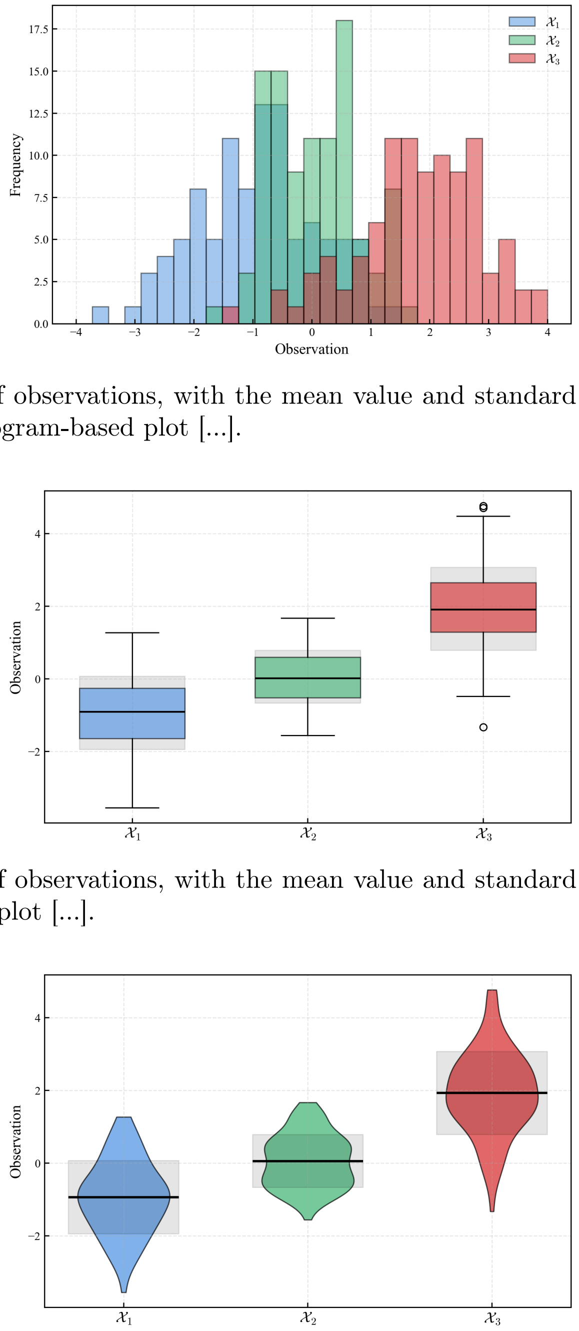
Figure 1.3: Box plots representing $n = 100$ observations drawn from a Gaussian distribution. The central black line shows the mean value, representing the central tendency where the bulk of events lie. The shadowed area highlights the standard deviation, as measure of the variability and spread the observations with respect to the mean

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].



(a) Three sets of observations, with the mean value and standard deviation represented as a histogram-based plot [...].

(b) Three sets of observations, with the mean value and standard deviation represented as a box plot [...].

(c) Three sets of observations, with the mean value and standard deviation represented as a violin plot [...].

Figure 1.4: Comparison of three visualization methods—histogram, box plot, and violin plot—showing the mean and variability of three samples of size $n = 100$.

Chapter 2

Foundations of Probability

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

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

Beyond games of chance, probability rapidly became essential for the understanding of natural phenomena and human affairs. Astronomy, population studies, and various physical problems required tools to reason quantitatively about variability, error, and incomplete information. In this sense, probability emerged not merely as a mathematical curiosity, but as a response to practical problems involving uncertainty and regularity in the empirical world.

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

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

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

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

Neyman—focused on statistical inference from data rather than individual decision-making. These approaches emphasized long-run frequency properties, error control, and sampling distributions, leading to the classical framework of statistical inference that still underlies much of modern applied statistics.

2.1 Probability and random events

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

In modern mathematics, probability is defined axiomatically following Kolmogorov [15]. A probability $\mathbb{P}$ assigns a number to each event and satisfies three fundamental rules:

- Probabilities are never negative: $\mathbb{P}(A) \geq 0$ for any event A.
- The probability of a certain event is 1.
- If two events cannot occur together, the probability that one or the other occurs is the sum of their probabilities.

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

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

which simply states that *something must happen.*

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

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

One common interpretation of probability is the *frequentist* view, developed most clearly by von Mises [17]. In this perspective, probability is identified with the long-run relative frequency of an event in repeated, identical experiments. Saying that a coin has probability 0.5 of landing heads means that, over many tosses, roughly half will result in heads.

An alternative interpretation is the *Bayesian* view, originating with Bayes [18] and developed further by Laplace and later authors such as de Finetti [19] and Jaynes [20]. Here, probability quantifies uncertainty or degree of belief rather than long-run frequency. Probabilities are updated as new information becomes available, using Bayes' theorem.

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

2.2 Discrete events

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

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

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

2.2.1 Bernoulli trials

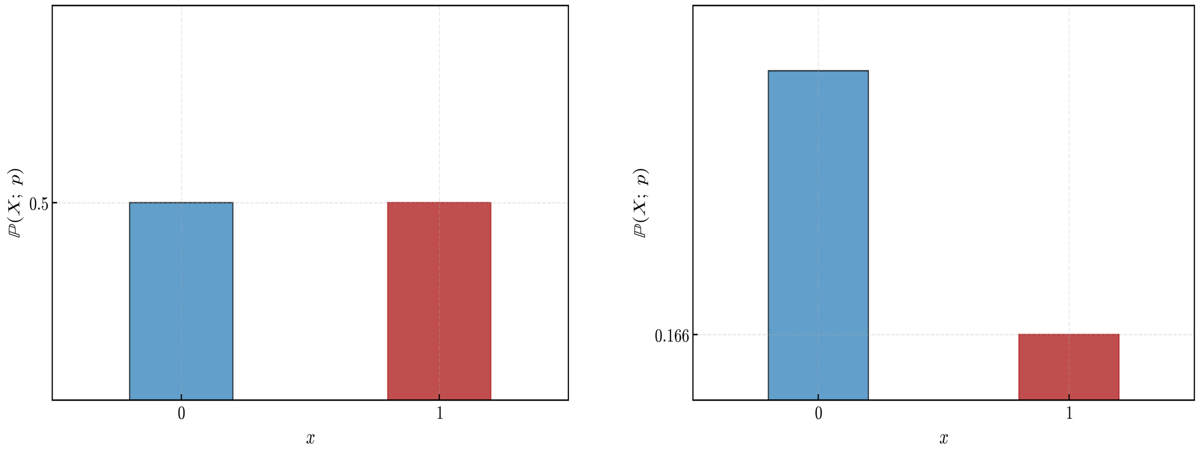
The Bernoulli trial was formalized by Jacob Bernoulli in *Ars Conjectandi* (1713) [21]. His motivation was to understand how regularity emerges from randomness when an experiment with two outcomes is repeated many times.

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

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

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

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



(a) Bernoulli example 1.

(b) Bernoulli example 2.

Figure 2.1: Probability distribuion

2.2.2 Binomial distribution

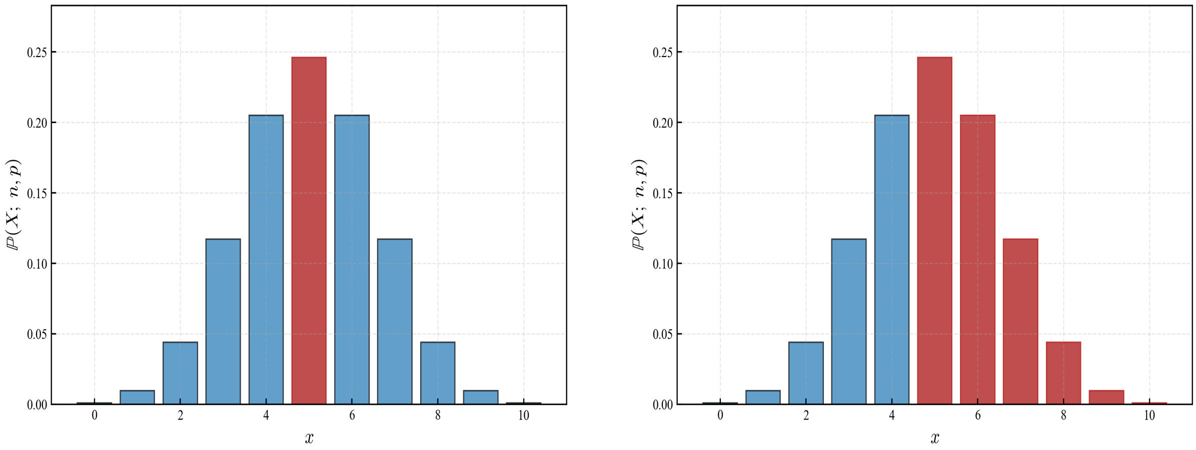
The binomial distribution was systematically studied by Abraham de Moivre in the early eighteenth century. His analysis of repeated Bernoulli trials led not only to the binomial formula but also to the first appearance of the normal approximation. De Moivre introduced the binomial distribution while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

The binomial distribution models the number of successes in n independent Bernoulli trials with success probability p. Its probability mass function is

$$\mathbb{P}(x;\ n,\ p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n. \tag{2.3}$$

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

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



(a) Probability distribution.

(b) Cumulative probability.

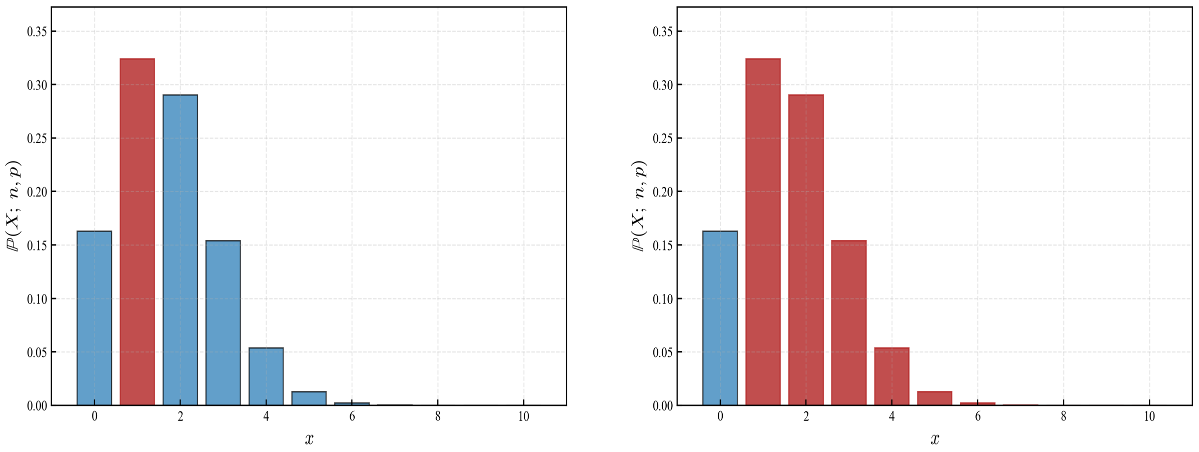
Figure 2.2: Probability distribiuon

2.2.3 Poisson distribution

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

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

$$\mathbb{P}(x;\ \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, 2, \ldots, \tag{2.4}$$



(a) Probability distribution.

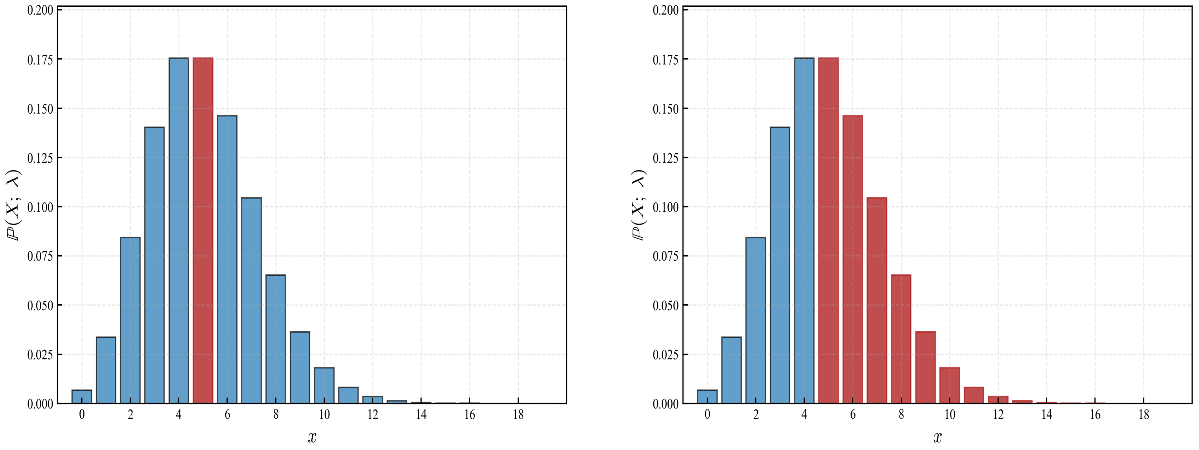
(b) Cumulative probability.

Figure 2.3: Cumulative probability

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

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

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



(a) Probability distribution.

(b) Cumulative probability.

Figure 2.4: Probability distribuion

2.2.4 Discrete uniform distribution

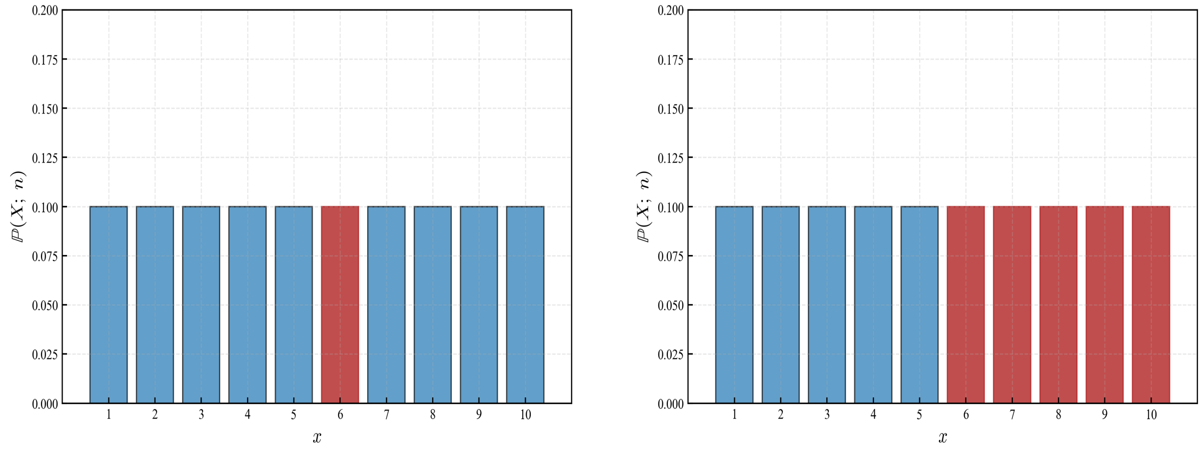
The discrete uniform distribution has its roots in classical symmetry arguments used in early probability theory. It formalizes the idea that, in the absence of distinguishing information, all outcomes should be treated equally. This idea reflects a principle already present in early probability theory: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

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

$$\mathbb{P}(x_i;\, n) = \frac{1}{n}, \qquad i = 1, \ldots, n. \tag{2.5}$$

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

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



(a) Probability distribution.

(b) Cumulative probability.

Figure 2.5: Probability distribuion

2.3 Continuous events

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

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

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

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

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

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

2.3.1 Gaussian distribution

The Gaussian distribution emerged from the work of Abraham de Moivre in the early eighteenth century and was later developed systematically by Pierre-Simon Laplace. Its physical interpretation was provided by Carl Friedrich Gauss in *Theoria Motus Corporum Coelestium* (1809) [22], in the context of measurement errors. Gauss introduced the distribution while studying astronomical observations, where repeated measurements of the same quantity produced small deviations around a central value. This interpretation linked probability theory directly to experimental science.

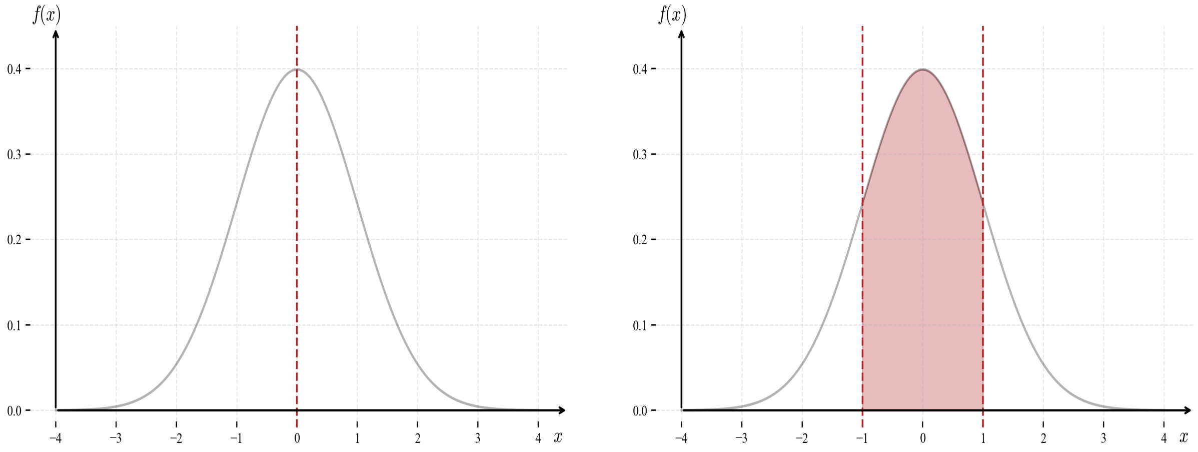
The Gaussian distribution models the accumulation of many small, independent effects. Its central role in probability theory is explained by the central limit theorem, which establishes it as a universal limiting distribution.

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

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

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

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



(a) Probability distribution.

(b) Cumulative probability.

Figure 2.6: Probability distribuion

2.3.2 Exponential distribution

The exponential distribution arose in the nineteenth century in the study of waiting times and decay processes, closely connected to Poisson's work on random events and later developments in queueing theory [23, 24]. The exponential distribution emerged naturally from the study of random event timing, particularly in physics and telecommunications. Its mathematical form reflects the assumption that events occur independently and at a constant average rate.

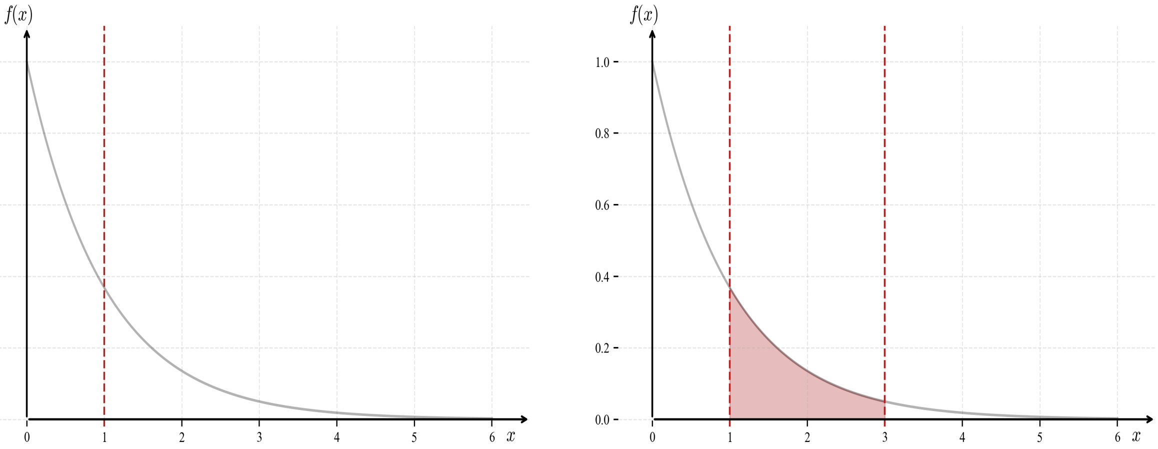
It naturally models the time until the first occurrence of a random event and is characterized by the absence of memory: the future waiting time does not depend on how much time has already elapsed.

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

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

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

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



(a) Probability distribution.

(b) Cumulative probability.

Figure 2.7: Probability distribuion

2.3.3 Continuous uniform distribution

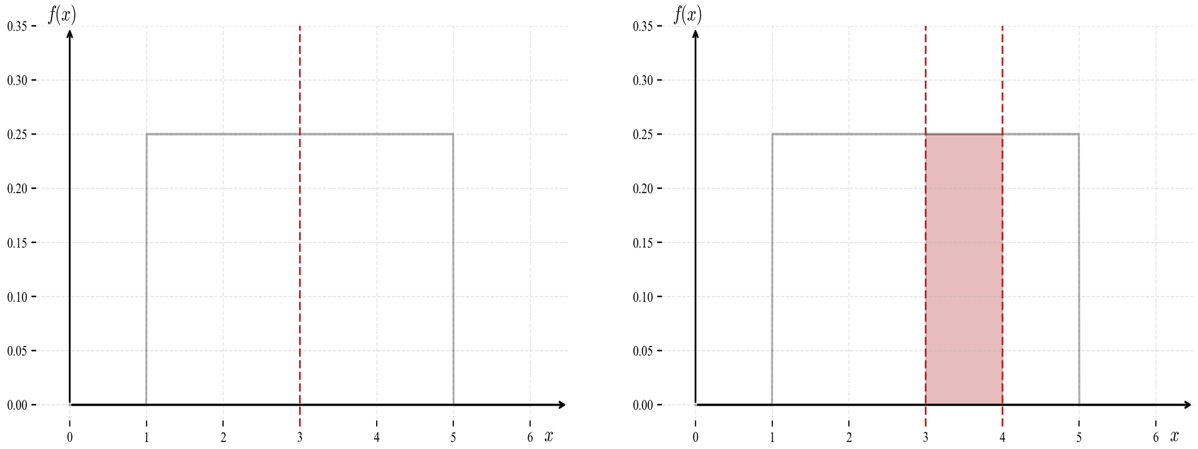
The continuous uniform distribution extends classical symmetry arguments already present in Laplace's *Théorie Analytique des Probabilités* (1812) [25]. It represents complete ignorance about where within a bounded interval an outcome may fall. The continuous uniform distribution formalizes the idea of complete uncertainty over a bounded range. Unlike other distributions, it does not privilege any value within the interval, making it a neutral reference model.

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

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

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

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



(a) Probability distribution.

(b) Cumulative probability.

Figure 2.8: Probability distribuion

2.4 Expected values

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

The most fundamental moment is the *expected value*, also called the mean. Informally, the expected value represents the long-run average outcome of a random experiment repeated many times. It answers the question: *where is the distribution centered?*

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

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

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

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

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

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

regardless of whether the variables are independent.

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

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

An equivalent and often more convenient expression for the variance is

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

For a discrete random variable, this corresponds to

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

while for a continuous random variable it is given by

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

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

- The *first moment* (mean) describes location.
- The *second moment* (variance) describes spread.
- The *third moment* is related to *skewness*, measuring asymmetry.
- The *fourth moment* is related to *kurtosis*, measuring tail heaviness.

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

2.5 Mean and variance of common distributions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Chapter 3

Estimation, variability and confidence

Numbers have an important story to tell, if given a voice.

— Florence Nightingale

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

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

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

3.1 The Law of Large Numbers

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

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

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

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

converges to μ as the sample size n increases.

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

3.2 The Central Limit Theorem

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

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

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

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

approaches a standard normal distribution as n becomes large.

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

3.3 Bias, variance and Mean Squared Error

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

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

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

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

and its variance is

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

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

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

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

3.4 Confidence intervals and critical regions

Point estimates summarize data with a single number, but they do not convey uncertainty. Confidence intervals address this limitation by providing a range of plausible values for an unknown parameter.

A confidence interval is constructed so that, in repeated sampling, it contains the true parameter a fixed proportion of the time. This proportion is called the confidence level and is typically expressed as a percentage.

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

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

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

Confidence intervals are closely related to hypothesis testing. The set of parameter values not rejected by a statistical test forms a confidence region. This duality provides a unified framework for estimation and decision-making.

Example. A 95% confidence interval for the mean expresses the range of values that are consistent with the observed data under repeated sampling.

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Chapter 4

Introduction to hypothesis testing

The object of statistical science is the reduction of data to relevant information.

— Ronald A. Fisher

The term *hypothesis testing* lies on top of the two pillars we have mentioned in previous chapters. From statistical analysis, as discussed in Chapter 1, we will use sampling, estimators, and graphical summaries to describe data. From probability theory, as discussed in Chapter 2, we will rely on probability distributions and expected values to model random variation and uncertainty. Through this chapter, we will mostly work in a *parametric* setting, where data is assumed to follow a simple, smooth and well-behaved distribution—most often the Gaussian distribution. Under these assumptions, and supported by the Law of Large Numbers and the Central Limit Theorem, estimators such as the sample mean and variance provide reliable information about the true population parameters. Confidence intervals and critical regions, introduced in Chapter 3, formalize this idea of reliability.

Within this framework, we introduce the notion of a *statistic*, or *statistic test*: a numerical quantity computed from data, designed to measure how close our observations are to what we would expect under a given hypothesis. Statistical tests are built by comparing such statistics to their behavior under an assumed model. In the modern Neyman–Pearson framework, this comparison is summarized through the *P-value*, which quantifies how unusual the observed data would be if the null hypothesis were true [26].

Let us begin with a brief historical note, often skipped. The ideas behind hypothesis testing did not emerge all at once. While the general notion of testing expectations against observations is very old, the mathematical tools required for modern statistical testing are relatively recent. Systematic use of estimators and probability models developed gradually during the nineteenth and early twentieth centuries, and the axiomatic foundations of probability were only formalized in 1933 with Kolmogorov's work [15]. The first statistical tests were developed in the early twentieth century by researchers such as Pearson and Fisher, among many other, and were originally designed to measure discrepancies between data and theoretical expectations, not to support automatic accept-reject decisions. The formal structure of hypothesis testing, including null and alternative hypotheses, error rates, and P-values, was later introduced by Neyman and Pearson. Because these ideas developed in stages, their interpretation requires some care. Throughout this chapter, we will emphasize both the practical use of statistical tests and the assumptions on which they are based.

4.1 Prediction vs inference revisted

When formulating hypotheses about natural phenomena, it is important to recall the distinction between population and sampling. On the one hand, we consider an idealized and generally inaccessible population, characterized by true but unknown quantities such as the mean and variance. Mathematical prediction, in this context, refers to the computation of expected values—also known as statistical moments—defined for a random variable together with its probability distribution.

Given a random variable, expected values (or moments) can be computed directly from its distribution.

Population mean for a discrete random variable x with support $\{x_i\}$

$$\mu = \mathbb{E}[x] = \sum_i x_i \, \mathbb{P}(x = x_i) \tag{4.1}$$

Population variance for a discrete random variable x

$$\sigma^2 = \mathbb{E}\big[(x-\mu)^2\big] = \sum_i (x_i - \mu)^2 \, \mathbb{P}(x = x_i) \tag{4.2}$$

Population mean for a continuous random variable x with density $f(x)$

$$\mu = \mathbb{E}[x] = \int_{-\infty}^{\infty} x \, f(x) \, dx \tag{4.3}$$

Population variance for a continuous random variable x

$$\sigma^2 = \mathbb{E}\big[(x-\mu)^2\big] = \int_{-\infty}^{\infty} (x-\mu)^2 \, f(x) \, dx \tag{4.4}$$

Hypotheses are therefore always formulated in terms of mathematical predictions about population-level parameters. For example, under Newtonian mechanics, a hypothesis may consist of Newton's second law, from which predictions can be made about the position of a falling object as a function of time. In biology, a hypothesis may concern the effect of a gene on a disease or on stress response, formulated in terms of expected expression levels or counts. In epidemiology, one may hypothesize a relationship between smoking prevalence and lung cancer risk. In all such cases, hypotheses concern population properties, while access to them is obtained only through finite samples of observations, collectively referred to as *data*.

Observed sample mean for a sample $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4.5}$$

Observed sample variance for the same sample

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4.6}$$

Hypothesis testing formalizes this comparison between population-level predictions and sample-based estimates by quantifying how compatible the observed data are with a hypothesized model.

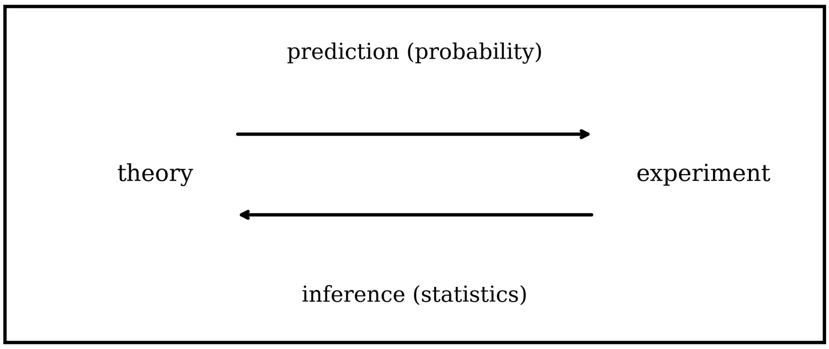


Figure 4.1: Representation of the predictive (from theory, or model, to experimental verification) and inferential (from data, measurement, observation to underlying truth) approaches to natural phenomena. As an example of the predictive branch of mathematics dealing with uncertainty we would find the theory of probability, while the descriptive way of addressing the same problem is normally regarded as statistical inference.

4.2 General approach to hypothesis testing

When dealing with hypotheses, predictions, experiments, and data, there exist many approaches and formulations, as many as instruments, scales, and fields of study. These approaches change from one field to another, and they also change over time. Our very ideas of hypothesis, prediction, measurement, and scientific law have evolved historically. Nowadays, when people refer to *hypothesis testing*, they usually mean a very specific approach: an almost algorithmic set of rules applied broadly to inference and data analysis problems. In this chapter, we will define such an approach as the *modern* or *general* approach to hypothesis testing. It assumes basic notions of probability theory, randomness, and probability distributions, together with statistical concepts such as estimators and sample-based descriptions. The core ideas of statistical tests, p-values, and significance that we discuss here are relatively recent, tracing back to the work of Pearson, Fisher, and Neyman in the early twentieth century.

The general approach to hypothesis testing can be summarized in the following steps:

- **Formulate hypotheses.** One specifies a *null hypothesis* H_0, representing the expected or reference case, and an *alternative hypothesis* H_1, representing a departure from H_0 that would be considered surprising. These hypotheses are always formulated in terms of true population parameters, and are often expressed through expected values or related quantities discussed in Chapter 2.

- **Experiment, measurement, observation.** Any process—regardless of instrumentation or field of study—that produces measurements or observations, resulting in one or more samples of data.

- **Compute a statistic or test statistic.** From the observed data, one computes an informative quantity, which may be a simple estimator such as the sample mean or variance, or a more elaborate statistic designed to quantify how far the observed data deviate from what is expected under H_0.

- **Compute a p-value.** The p-value is the probability that, assuming the null hypothesis and its associated population parameters are true, one would obtain a value of the statistic at least as extreme as the one observed.

- **Interpret the result.** The p-value is compared to a chosen significance level, leading to a decision or conclusion, commonly phrased as rejecting or not rejecting the null hypothesis.

A few remarks are in order regarding this general roadmap. A statistic may be as simple as an estimator, such as the sample mean, or a more abstract quantity derived from the data. Fisher originally introduced the p-value as a continuous measure of evidence against the null hypothesis, rather than as a strict decision rule. The widespread practice of fixed significance thresholds and accept–reject decisions reflects a later synthesis of Fisher's ideas with the Neyman–Pearson framework. As a result, the modern approach to hypothesis testing combines elements of both traditions, a point we will return to later in this chapter. *Warning.* A p-value does not measure the probability that a hypothesis is true or false, but rather how compatible the observed data are with the null hypothesis under the assumed model.

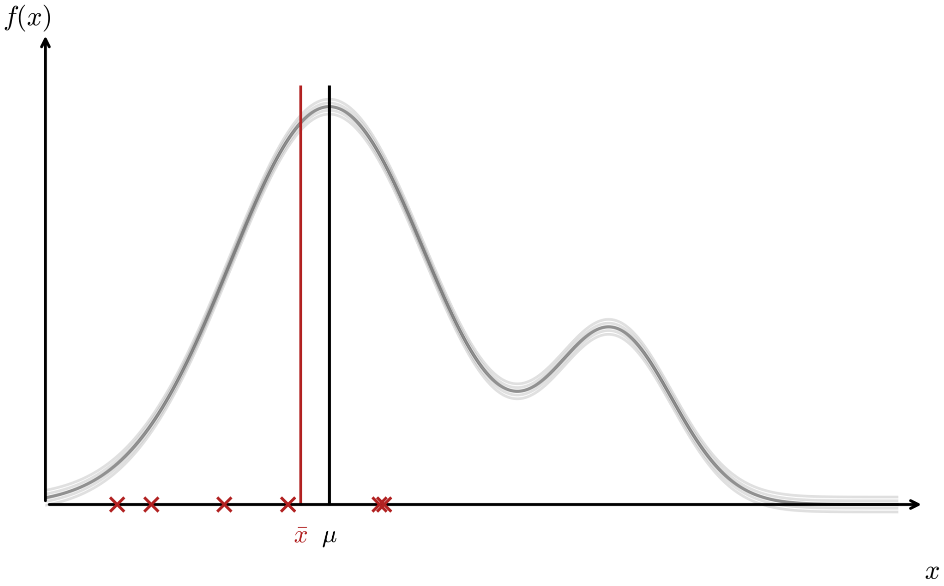


Figure 4.2: Representation of the *true* population mean μ, in black, and the observed *sample* mean $\bar{x}$. The true mean is and ideal and unaccessible quantity, while the sample mean can be computed as an estimator of the finite sample.

4.3 Statistic tests: common examples

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

The t-test is arguably the simplest example of a statistical test we will discuss. It was developed in 1908 by William S. Gosset, a statistician working at the Guinness factory in Dublin, who was trying to accurately estimate the error of the mean when the population variance is unknown, as part of the brewing process. Due to his affiliation with the Guinness company, he was not allowed to publicly share his work, and therefore submitted it to the journal *Biometrika* under the pseudonym *Student*. For this reason, it remains nowadays known as the *Student's t-test* [27].

The test begins by formulating a null hypothesis about the true population mean, *prior to any sampling or data collection*. Normally, the null hypothesis is written as *the true population mean is expected to take the value* μ. It is important here to stop and think carefully about what physical quantity we are actually going to measure. Recall that such a quantity, inaccessible in principle, can be either predicted through the computation of an expected value or estimated from data, as discussed in Chapter 2.

Now take a series of observations or measurements and group them into a sample $\chi = \{x_1, x_2, \ldots, x_n\}$. From this sample, we can compute the sample mean $\bar{x}$ as an estimator of the true population mean μ, and the sample standard deviation s as an estimator of the true population standard deviation σ.

Given these three elements (the expected value μ, given by our null hypothesis H_0, together with the sample mean $\bar{x}$ and the sample standard deviation s, obtained from data) we can compute the *t-statistic*, or *t-test statistic*, defined as

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \ . \tag{4.7}$$

Let us pause for a moment to examine this quantity. As the sample mean $\bar{x}$ approaches the expected value μ, the t-statistic tends to zero. It was designed precisely for this purpose: to quantify how different—or how similar—our observed data are with respect to the hypothesized value, so that in the ideal case $\bar{x} \to \mu$, we have $t \to 0$.

It is important to note that t is computed from a set of random observations. If we repeat the same measurements using a different sample, or at a different time, or under different conditions, we may obtain different values of $\bar{x}$ and s, and therefore a different value of t. This means that *t is itself a real-valued random variable*, and it follows some probability distribution. The mathematical form of this distribution, introduced in Gosset's original work, is known as the *Student's t-distribution*,

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

where the parameter ν is referred to as the *degrees of freedom* and is simply related to the sample size by $\nu = n - 1$. Some textbooks write the statistic as t_ν, which is standard notation, but here we reserve the symbol t for the statistic itself and $f(t;\nu)$ for its distribution. It can be shown that as the degrees of freedom increase, the t-distribution converges to the standard normal distribution, written as $f(t;\nu) \to \mathcal{N}(0,1)$ as $\nu \to \infty$; the explicit proof is beyond the scope of this course.

We now arrive at the final step: the computation of the p-value. Returning to Fisher's definition of the p-value as *the probability of obtaining a value at least as extreme as the one observed for the statistic, assuming the null hypothesis H_0 is true*, we see that once the distribution of the t-statistic is known, the p-value can be computed by evaluating an appropriate cumulative probability.

If we ask for the probability of obtaining a value strictly greater (or strictly smaller) than the observed statistic, $\mathbb{P}(t \geq t_{\text{obs}})$, we compute the integral of one tail of the distribution; this is called a *one-sided* (or one-tailed) p-value. If instead we ask for the probability of obtaining a value more extreme than the observed one, regardless of direction, $\mathbb{P}(|t| \geq |t_{\text{obs}}|)$, we integrate both tails of the distribution; this is called a *two-sided* (or two-tailed) p-value. Due to the symmetry of the t-distribution, the two-sided p-value is simply twice the one-sided p-value.

$$p_{\text{one-sided}} = \mathbb{P}(t \geq t_{\text{obs}}) = \int_{t_{\text{obs}}}^{\infty} f(t;\nu)\,dt \ ,$$

$$p_{\text{two-sided}} = \mathbb{P}(|t| \geq |t_{\text{obs}}|) = 2\int_{|t_{\text{obs}}|}^{\infty} f(t;\nu)\,dt \ .$$

For a review of cumulative probabilities and probability integrals, see Chapter 2, and for a brief refresher on integral calculus, consult Appendix C. In practice, numerical computation of p-values is typically carried out using statistical software, which evaluates both the test statistic and the corresponding tail probabilities; common examples include the `scipy` library in `Python` and the `stats` package in `R`.

Example. For a sample of size $n = 10$, observed mean $\bar{x} = 5.2$, sample standard deviation $s = 1.0$, and hypothesized value $\mu = 5$, we obtain

$$t_{\text{obs}} = \frac{5.2 - 5}{1/\sqrt{10}} \approx 0.63.$$

With $\nu = 9$ degrees of freedom, the corresponding two-sided p-value is approximately $p \approx 0.54$.

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

The two-sample t-test is an extension of the one-sample case, hence the general approach will be almost identical as the one discussed in previous section. It is used to test two independent samples, chi_1 and χ_2, of lengths n_1 and n_2. The null hypothesis is still formulated about the true population means, as *both observations come—are sampled from—the same distribution, with an expected true mean μ.* Remember again that such quantity, unaccessible in theory, can be either fitted from previous data, or predicted through the computation of an expected value, as we discussed in Chapter 2.

Now take a series of measurements for both samples, and compute the sample means $\bar{x}_1$, $\bar{x}_2$, as estimators of the true population mea μ, and the sample variances s_1^2, s_2^2, as estimators of the true population variance σ^2.

In the same way we proceeded for the one-sample case, we combine the expectation given by our H_0 and the estimators computed out of our data, into the *two-sample t-statistic*, or *two-sample t-statistic test*, defined as

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

The denominator is normally called *pooled standard deviation*, and denoted by s_p. We can see that this quantity behaves in the same way as the one-sample case, tending to zero, as the sample means of both groups tend to each other, $t \to 0$ as $\bar{x}_1 \to \bar{x}_2$,

Again, t is a real-valued random variable, and it follows the same Student's t-distribution of eq. (4.8). The only difference is that the degrees of freedom ν are now obtained by combining the lengths of both samples $\nu = n_1 + n_2 - 2$.

The computation of the P-value, is identical to the one-sample case, as we just need to integrate the t-distribution. Again, given the symmetry of the t-distribution, the two-sided P-value is just double the size of the one-sided case. For a review of cumulative probabilities and integrating probability distributions, go back to Chapter 2, and for a review on integral calculus and some warm-up examples, see Appendix C.

Example. For sample of size $n_1 = 10$, observed average $\bar{x} = 5.2$, variance $s = 1.0$, and $\mu_0 = 5$, then

$$t_{\text{obs}} = \frac{5.2 - 5}{1/\sqrt{10}} \approx 0.63,$$

Given this observed value, and the degrees of freedom $\nu = 9$, the two-sided p-value would be $p \approx 0.54$.

4.3.3 Fisher's F test: Compare variation of two independent groups

The Fisher's variance-ratio test, or F-test for short, was introduced by Fisher in the 1920s and formally developed in his works *Statistical methods for research workers* and *The design of experiments*, in 1925 and 1935. The impact of Fisher's work, not only in statistics but also in evolutionary biology, experimental design, hypothesis tesing and mathematical modelling is credited still nowadays as one of the greatest among the twentieth century, by far [28, 29, 30].

The F test begins by formulating some null hypothesis about the true population variance, *prior to any sampling or data collection.* Normally, the null hypothesis is simply written as *both samples under study come from the same distribution, with true population variance* σ^2. As we did in previous cases, remember that such quantity is hypothesized value about the true population, and it can be either fitted from data, or predicted through the computation of an expected value, as we discussed in Chapter 2.

Now take a series of measurements for two independent samples χ_1, chi_2 of sizes n_1 and n_2, compute the sample variances s_1^2, s_2^2, as estimators of the true population variances σ_1^2 and σ_2^2. Then the Fisher F-*statistic*, or F-*statistic test*, is defined just as the ratio

$$F = \frac{s_1^2}{s_2^2} \, . \tag{4.10}$$

Similarly to what happend in previous cases, we can notice that as the sample variances s_1^2, s_2^2 approach each other, the F-variable tends to one. It was designed for this precise purpose, to quantify how different—or similar—two independent groups are from each other, and in the ideal case $s_1^2 \to s_2^2$, then $F \to 1$. Recall the definition of the t-statistic, as here we can start noticing that, in general, statistic tests are normally defined such that, in the case of H_0 being true, they reduce to a small, simple value.

If we pay close attention, we can notice that unlike the t-test, where the null hypothesis was stated directly in terms of a μ parameter, appearing explicitly in the definition of the t-statistic, there is no explicit trace of H_0 in the definition of our F, which is computed just as the ratio of two sample variances. There is a historical reason for this, that we will revisit further in this chapter, related to how the very idea of hypotheis, parameter and statistic test were used in Fisher's time, different from modern usage. As we will see, the classical F-test encodes the null hypothesis through the *condition under which the ratio of F of sample variances follow a specific distribution.* In practice, this reflects the fact that the t-test is formulated around an explicit parameter hypothesis, whereas the F-test arose from Fisher's analysis of sampling distributions.

Same as in the two previous examples, the F-statistic is obtained out of a set of random observations. If I repeat the same measurements in a different sample, or a different day, or under different conditions, they may lead to different values s_1 and s_2, hence producing a different F. This means that F *is a real-valued random variable itself*, and it will follow *some* distribution. The mathematical definition of such distribution—or density—of the F-variable, as introduced in Fisher's [...], is called the Fisher's F-*distribution*,

$$f(F;\nu_1,\nu_2) = \frac{1}{B\!\left(\dfrac{\nu_1}{2},\dfrac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} F^{\frac{\nu_1}{2}-1} \left(1+\frac{\nu_1}{\nu_2}F\right)^{-\frac{(\nu_1+\nu_2)}{2}} \quad , \tag{4.11}$$

where the parameters ν_1, ν_2 represent the *degrees of freedom*, and they are related to the length of the samples $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$. You may see that some textbooks and literature sources write it as F_{ν_1,ν_2}, which is perfectly fine and common standard, but we prefere to use here the letter F just for the statistic, and $f(F;\nu_1,\nu_2)$ for the distribution of F given ν_1 and ν_2 degrees of freedom, rather than referring to both elements with the same symbol. It can be demonstrated that as $\nu_1, \nu_2 \to \infty$, $f(t; \nu_1, \nu_2)$ concentrates at 1 and $\log f(F; \nu_1, \nu_2)$ becomes approximately Gaussian. The explicit demonstration is out of the scope of this course, but it can be found at [...].

As we mentioned alread, it is under the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ that the F-statistic follows the Fisher distribution. Modern Wald and likelihood-ratio tests provide a unified parameter-based framework.

Back to the computation of the P-value, given Fisher's definition, *the probability of obtaining a value at least as extreme as the one observed for our statistic, asuming the expected value—or values—given by our null hypothesis H_0 and our random data*, we just need to compute the cumulative probability—that is, the integral—of the F distribution.

If we ask for the probability of obtaining a value *strictly greater or strictly smaller than the one we observed for our statistic*, $\mathbb{P}(F \geq F_{\text{obs}})$, we just need to compute the integral of one of the distribution tales, and it give the one-sided P-value. But the two-sided case, the probability of obtaining a value *more extreme, regardless of the direction* would require the integral of both tails, and given the asymmetry of the F-distribution, becomes a non-trivial task. The common formulation is written as follows

$$P_{\text{one-sided}} = \mathbb{P}(F \geq F_{\text{obs}}) = \int_{F_{\text{obs}}}^{\infty} f(F; \nu_1, \nu_2)\, df \ ,$$

$$P_{\text{two-sided}} = 2 \min \left\{ \int_{0}^{F_{\text{obs}}} f(F; \nu_1, \nu_2)\, df,\ \int_{F_{\text{obs}}}^{\infty} f(F; \nu_1, \nu_2)\, df \right\} .$$

For a review of cumulative probabilities and integrating probability distributions, go back to Chapter 2, and for a review on integral calculus and some warm-up examples, see Appendix C. If this feels a bit heavy, this is where computer softwares and libraries become particularly useful, as they not only implement the calculation of the statistic but the numerical integration of such distribution, yealding to the P-value without the need for manual integral calculus. Examples of these will be `scipy` library of `Python`, and the `stats` package of `R`, among many others [...].

Example. If $n_1 = n_2 = 10$, $S_1^2 = 4$, $S_2^2 = 2$, then

$$f_{\text{obs}} = 2, \qquad \nu_1 = \nu_2 = 9,$$

yielding a one-sided p-value $p \approx 0.12$.

4.3.4 Fisher's ANOVA: Compare variation of multiple groups

In the same way the two-sample t-test is an extension of the one-sample case, Fisher's analysis of variance (ANOVA) can be seen as a natural extension of the variance-ratio F-test to more than two groups. It was developed by Fisher in the 1920s as part of his investigation of sampling distributions under normality and of experimental design, aiming to assess whether several sources of variability could plausibly be attributed to the same underlying population variance [29].

In short, ANOVA tests whether the variability between group means is large relative to the variability within groups.

Now consider k independent samples $\chi_1, \chi_2, \ldots, \chi_k$, with sizes $n_1, n_2, \ldots, n_k$, sample means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, and an overall mean $\bar{x}$. The null hypothesis is formulated as *all groups come from populations with the same true mean*, that is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.

The ANOVA F-statistic is defined as the ratio of the variance between groups to the variance within groups,

$$F = \frac{s^2_{\text{between}}}{s^2_{\text{within}}} \ , \tag{4.12}$$

where

$$s^2_{\text{between}} = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2, \qquad s^2_{\text{within}} = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2,$$

and $N = \sum_{i=1}^{k} n_i$. As in the previous cases, the statistic is constructed so that, when the null hypothesis is true and group means are similar, the ratio remains close to one.

The F-statistic is again a real-valued random variable, and under the null hypothesis it follows a Fisher F-distribution with degrees of freedom $\nu_1 = k-1$ and $\nu_2 = N-k$. P-values are computed by integrating the right tail of this distribution, exactly as in the two-sample F-test.

4.3.5 Pearson's χ^2 test: Compare distributions and testing for normality

The chi-square test was introduced by Karl Pearson in 1900 as part of his work on goodness-of-fit and contingency tables [31]. Pearson's original formulation was oriented toward measuring discrepancy between observed and expected frequencies, rather than toward formal decision-based hypothesis testing. Given observed counts $\{O_1, O_2, \ldots, O_k\}$ and expected counts $\{E_1, E_2, \ldots, E_k\}$, the χ^2-statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{4.13}$$

was conceived as a numerical measure of lack of agreement between data and model, with larger values indicating poorer fit.[1]

As in the previous examples, the χ^2-statistic is computed from random data, and repeating the experiment would generally lead to different observed counts and hence a different value of χ^2. This means that χ^2 *is itself a random variable*, and it follows a probability distribution known as the Pearson χ^2-distribution,

$$f(\chi^2;\nu) = \frac{(\chi^2)^{\nu/2-1}e^{-\chi^2/2}}{2^{\nu/2}\Gamma(\nu/2)}\ , \tag{4.14}$$

defined for $\chi^2 > 0$.

Unlike the t-test, where the degrees of freedom are always $n-1$, the degrees of freedom ν of the χ^2-test depend on the structure of the data and on how many parameters are estimated under the null hypothesis. In particular:

- For a goodness-of-fit test with k categories, $\nu = k - 1 - p$, where p is the number of parameters estimated from the data.
- For a test of independence in an $r \times c$ contingency table, $\nu = (r-1)(c-1)$.

In the modern framework, the chi-square test is formulated with an explicit null hypothesis and a p-value. One specifies a null hypothesis H_0 describing the expected distribution or independence structure, derives the asymptotic distribution $\chi^2 \sim \chi^2_\nu$ under H_0, and computes the p-value as the right-tail probability

$$p = \mathbb{P}(\chi^2 \geq \chi^2_{\text{obs}}) = \int_{\chi^2_{\text{obs}}}^{\infty} f(\chi^2;\nu)\, d\chi^2.$$

A deeper theoretical connection appears through likelihood-based testing: Pearson's χ^2-statistic is asymptotically equivalent to the likelihood-ratio statistic $-2\log\Lambda$, a result formalized by Wilks in the 1930s. Both converge to a chi-square distribution under the null hypothesis, providing a unifying asymptotic framework together with Wald and score tests.

[1] Pearson's original formulation predates the modern null/alternative hypothesis framework and the notion of Type I and Type II errors.

4.3.6 The Wald test: asymptotic behavior

The Wald test was developed by Abraham Wald in the 1930s as part of a general asymptotic theory of hypothesis testing for parametric models with large samples [32]. Its purpose is to test hypotheses about finite-dimensional parameters in statistical models using large-sample approximations. Unlike classical tests tailored to specific settings, the Wald test provides a general framework applicable to linear models, generalized linear models, and maximum likelihood estimation.

Conceptually, the Wald test generalizes classical parametric tests such as the t- and F-tests by formulating hypotheses directly in terms of model parameters. In this sense, it unifies earlier procedures within a common asymptotic theory.

Let $\hat{\theta}$ be an estimator of a parameter θ. The Wald statistic is defined as

$$W = (\hat{\theta} - \theta_0)^\top \widehat{\mathrm{Var}}(\hat{\theta})^{-1}(\hat{\theta} - \theta_0).$$

Under the null hypothesis and suitable regularity conditions,

$$W \xrightarrow{d} \chi^2_p,$$

where p is the number of tested constraints. P-values are computed from the upper tail of the chi-square distribution.

4.4 Parametric and non-parametric tests

Statistical tests are often described as *parametric* or *non-parametric*, a distinction that reflects both historical development and underlying philosophical views about statistical modeling. Parametric tests were developed first, at the beginning of the twentieth century, in a context where probability models were seen as idealized descriptions of data. These tests assume that observations come from a distribution belonging to a family described by a small number of parameters, such as the mean and variance. Statistical inference then focuses directly on these parameters.

In practice, parametric tests are frequently associated with the Gaussian (normal) distribution. This historical association arises because many of the foundational procedures of classical statistics—such as the t-test, the F-test, and analysis of variance—are exact under normality. As a result, the parametric versus non-parametric distinction is often informally described as "Gaussian versus non-Gaussian." This simplification is useful pedagogically, but it should be remembered that parametric models also include many non-Gaussian distributions, such as the binomial or Poisson.

Non-parametric tests emerged later, largely in the 1940s and 1950s, motivated by the recognition that real data often deviate from idealized models. Rather than assuming a specific distributional form, these methods aim to remain valid under broad and unspecified distributions. They typically rely on ranks, signs, or empirical distributions, and therefore make fewer assumptions about the shape of the data.

From a historical perspective, the development of statistical testing can be roughly organized as a timeline. Early work by Pearson and Fisher between 1900 and 1930 established the foundations of parametric inference, including the chi-square test, the t-test, and analysis of variance. In the mid-twentieth century, researchers such as Wilcoxon, Mann, and Whitney introduced distribution-free methods to address practical limitations of these classical procedures. Later developments, including asymptotic theory and robust statistics, provided a unifying framework that connects parametric and non-parametric approaches.

To organize the tests presented in this section, it is helpful to focus on their *goal* rather than on their label. Some tests are designed to compare central tendencies between samples, others to assess variability, and others to evaluate the overall agreement between data and a theoretical model. Viewed this way, non-parametric tests are not simply substitutes for parametric ones, but complementary tools that reflect different assumptions, historical traditions, and inferential aims.

4.4.1 Wilcoxon signed-rank test

The Wilcoxon signed-rank test was introduced by Frank Wilcoxon in 1945 as a distribution-free alternative to the one-sample and paired-sample t-tests [33]. It was motivated by situations in which normality could not be assumed but a test of central location was still desired. The parametric analogue is the one-sample or paired t-test, which tests a hypothesis about a mean. By contrast, the Wilcoxon signed-rank test targets symmetry of the distribution about a specified location.

The test statistic is based on the ranks of the absolute deviations $|x_i - \theta_0|$, with signs retained. Under the null hypothesis of symmetry, the statistic has a known finite-sample distribution; for moderate to large samples, it is commonly approximated by a normal distribution, from which p-values are obtained.

4.4.2 Mann–Whitney U test

The Mann–Whitney U test, introduced by Mann and Whitney in 1947, provides a non-parametric alternative to the two-sample t-test for independent samples [34]. It was designed to compare two populations without assuming normality or equal variances. The parametric analogue is the two-sample t-test, which compares population means. The Mann–Whitney test instead assesses whether one distribution tends to produce larger observations than the other.

The test statistic U is constructed from the ranks of the pooled samples. Under the null hypothesis that the two distributions are identical, U has a known exact distribution and, asymptotically, a normal distribution. P-values are computed either exactly or via the normal approximation.

4.4.3 Levene median-based test

Levene's test was proposed by Howard Levene in 1960 as a robust alternative to Fisher's variance-ratio test [35]. The median-based version, later emphasized by Brown and Forsythe, improves robustness against non-normality. The parametric analogue is the classical F-test for equality of variances, which is highly sensitive to departures from normality. Levene's test replaces variances by absolute deviations from group centers.

The statistic is computed by applying a one-way ANOVA to the transformed data $|x_{ij} - \tilde{x}_i|$, where $\tilde{x}_i$ is the group median. Under the null hypothesis of equal spreads, the test statistic follows approximately an F distribution, from which p-values are obtained.

4.4.4 Kruskal–Wallis test

The Kruskal–Wallis test was introduced in 1952 by Kruskal and Wallis as a non-parametric extension of one-way ANOVA. It was motivated by the need to compare more than two groups without assuming normality. The parametric analogue is Fisher's one-way ANOVA, which tests equality of group means. The Kruskal–Wallis test instead evaluates whether the group distributions are identical [36].

The test statistic is based on the ranks of all observations:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i(\bar{R}_i - \bar{R})^2 \,.$$

Under the null hypothesis, H converges in distribution to χ^2_{k-1}. P-values are computed from the chi-square distribution.

4.4.5 The Kolmogorov–Smirnov test

The Kolmogorov–Smirnov test was developed in the 1930s by Kolmogorov and later extended by Smirnov as a general goodness-of-fit procedure [37, 38]. It compares an empirical distribution to a fully specified theoretical distribution. The parametric analogue is the chi-square goodness-of-fit test, which relies on binning and asymptotic approximations. The Kolmogorov–Smirnov test instead measures the maximum discrepancy between distribution functions.

The test statistic is

$$D = \sup_x |F_n(x) - F_0(x)| \,.$$

Under the null hypothesis, D has a known distribution independent of F_0. P-values are computed from this distribution or its asymptotic form.

4.4.6 The Shapiro–Wilk test

The Shapiro–Wilk test was introduced by Shapiro and Wilk in 1965 as a powerful goodness-of-fit test specifically designed to assess normality [39]. It was motivated by the low power of general-purpose tests when applied to normal models. The parametric analogue is not a test of means or variances, but rather the assumption of normality underlying t-tests, F-tests, and ANOVA. The Shapiro–Wilk test directly targets this assumption.

The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \,,$$

where the coefficients a_i depend on normal order statistics. The distribution of W under the null hypothesis is obtained via approximation or simulation, and p-values are computed accordingly.

4.5 Error types in hypothesis testing

Modern hypothesis testing emerged in the early twentieth century as an attempt to formalize uncertainty, error, and decision making in empirical science. Three major approaches—Fisherian significance testing, Neyman–Pearson hypothesis testing, and Bayesian inference—address these issues in fundamentally different ways, while later philosophical analyses by Reichenbach and Popper clarified their distinct aims. Subsequent commentators such as Cox, Mayo, and Lehmann have emphasized both the strengths of each framework and the conceptual tensions created by their later amalgamation in textbook practice.

Ronald A. Fisher introduced significance tests in the 1920s as tools for assessing the *strength of evidence* against a null hypothesis [29, 30]. In Fisher's view, a null hypothesis H_0 is a reference model, and the p-value is defined as the probability, under H_0, of observing data at least as extreme as those obtained. Small p-values indicate discordance between data and model, but Fisher rejected fixed decision thresholds and did not formalize Type II errors or power. Type I error appears implicitly as the tail probability under H_0, not as a long-run operating characteristic. Hypothesis testing, for Fisher, is evidential rather than decisional: it informs scientific judgment but does not prescribe action.

Jerzy Neyman and Egon Pearson developed a sharply different framework in the 1930s, motivated by repeated decision making [40]. Here, hypotheses H_0 and H_1 are competing models, and tests are designed to control error rates in the long run. Type I error (α) and Type II error (β) are central primitives, and optimal tests maximize power subject to a fixed α. P-values play no essential role; instead, decisions are based on pre-specified critical regions. This approach interprets hypothesis testing as a rule for action under uncertainty rather than as a measure of evidential support.

Bayesian inference, originating in Bayes's posthumous essay [18] and developed by Laplace and later subjectivists such as de Finetti [19], rejects Type I and Type II errors as fundamental concepts. Probability is interpreted as rational degree of belief, and hypotheses themselves are assigned probabilities. Inference proceeds by updating prior beliefs via Bayes' theorem to obtain posterior probabilities or Bayes factors. Hypothesis testing becomes model comparison, and decisions—if required—are made by minimizing expected loss. The Bayesian framework thus dissolves the classical error dichotomy by reframing uncertainty epistemically rather than behaviorally.

Hans Reichenbach provided the clearest philosophical articulation of the frequentist stance underlying Neyman–Pearson theory [41]. He distinguished *prediction*—statements about long-run frequencies—from *inference*—claims about truth or belief. Statistical tests, on this view, justify actions and predictions through their error properties, not through probabilistic assertions about hypotheses. This position sharply contrasts with Bayesian epistemology and clarifies why frequentist testing can function without assigning probabilities to hypotheses.

Karl Popper rejected probabilistic confirmation altogether, arguing that science advances through bold conjectures and severe attempts at falsification [42]. Statistical tests, in his view, contribute by formulating risky predictions whose failure can refute theories, not by accumulating evidence or controlling long-run errors. Popper's philosophy is incompatible with Bayesian confirmation and only partially aligned with frequentist testing, insofar as both emphasize error and refutation rather than belief.

Erich Lehmann, in his definitive treatment of hypothesis testing [43], emphasized the formal coherence and optimality of Neyman–Pearson theory while explicitly distinguishing it from Fisher's evidential approach. D. R. Cox later argued that the routine combination of p-values with fixed significance thresholds conflates logically distinct inferential goals [44]. Deborah Mayo further developed an error-statistical philosophy in which evidential interpretation is grounded in the severity with which hypotheses are tested [45, 46]. Together, these authors converge on a common diagnosis: the modern textbook procedure of hypothesis testing is a pragmatic but conceptually hybrid construct, blending incompatible foundations.

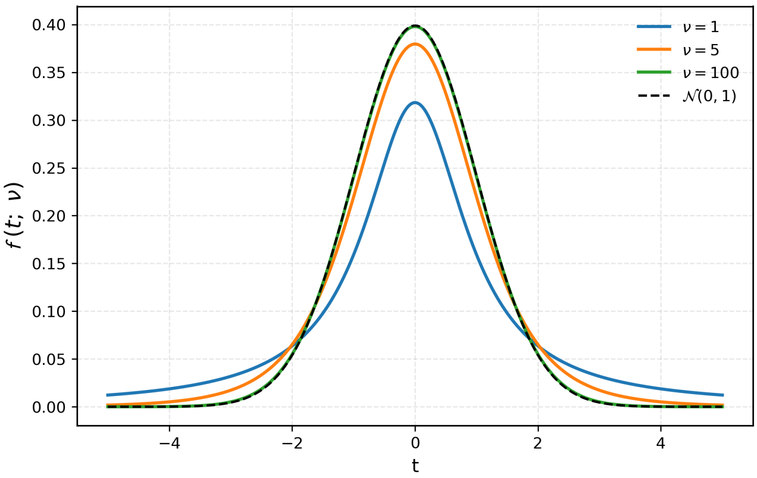
The coexistence of Fisherian evidence, Neyman–Pearson decision rules, Bayesian belief updating, Popperian falsification, and Reichenbach's predictive frequentism reflects not confusion but plurality. Each framework answers a different question—about evidence, action, belief, or prediction—and Type I and Type II errors acquire meaning only within the Neyman–Pearson decision-theoretic context. Understanding these distinctions is essential for the principled use and interpretation of hypothesis tests in modern statistics.

Exercises

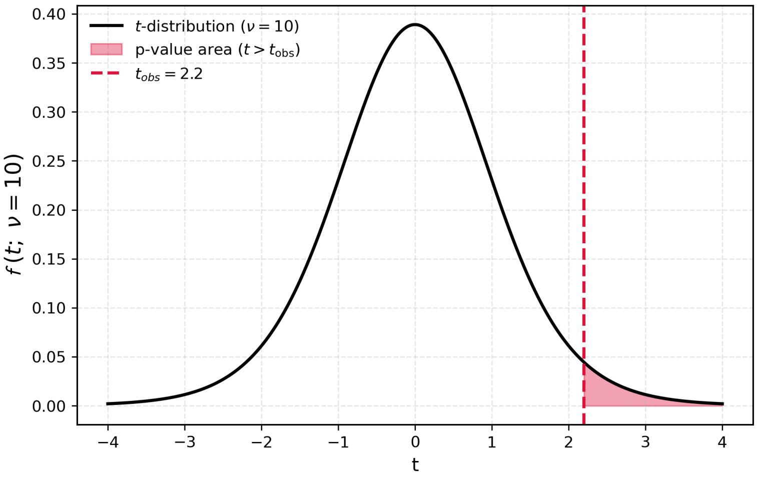
1. Exercise [...].

2. Exercise [...].

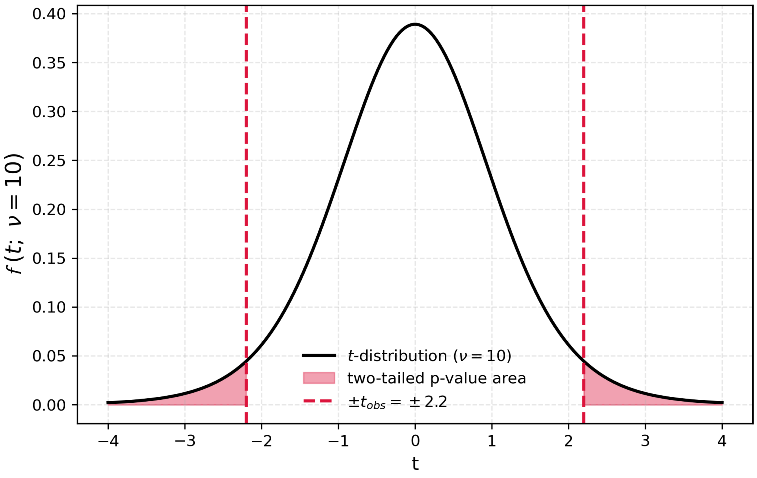
3. Exercise [...].



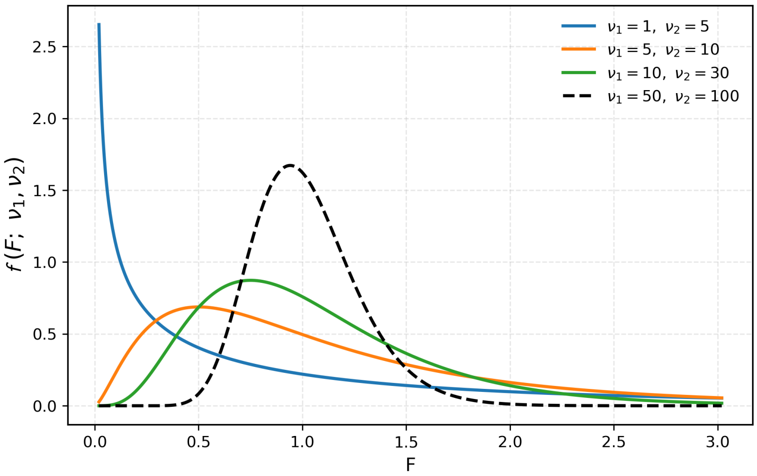
(a) The Student's t distribution of the t-statistic, given different values of the degrees of freedom ν.



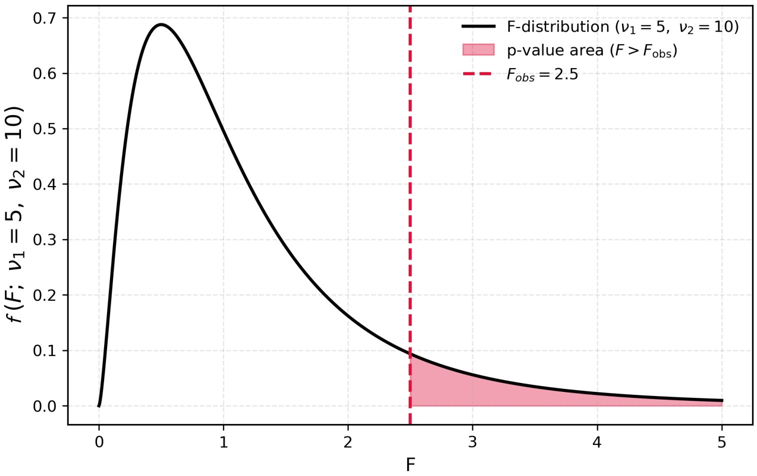
(b) Representation of the 1-sided P-value, computed as the integral of the right tail of the t distribution.



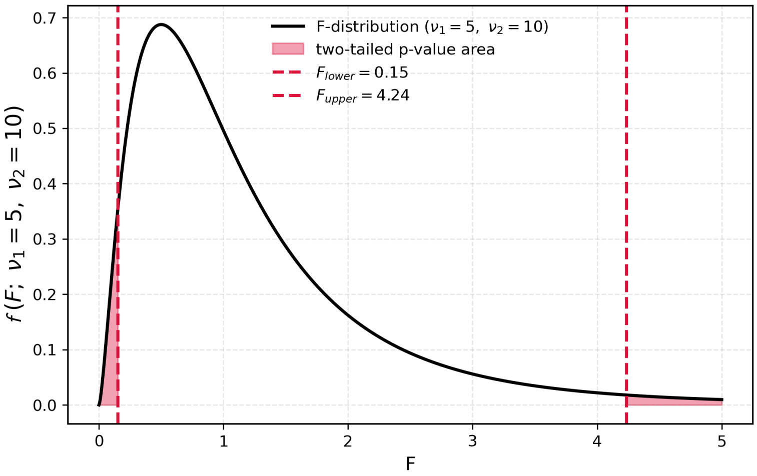
(c) Representation of the 2-sided P-value. Given the symmetry of the t-distribution, it can be obtained as double in size of the 1-sided integral.



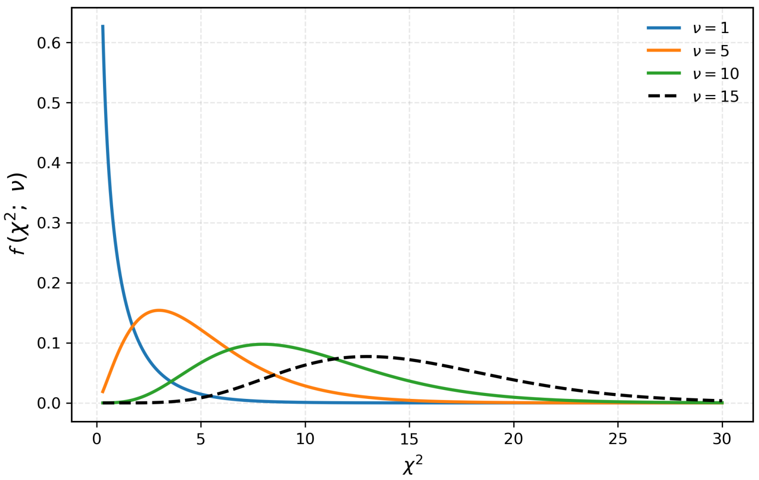
(a) The Fisher's F distribution of the F-statistic, given different values of the degrees of freedom ν.



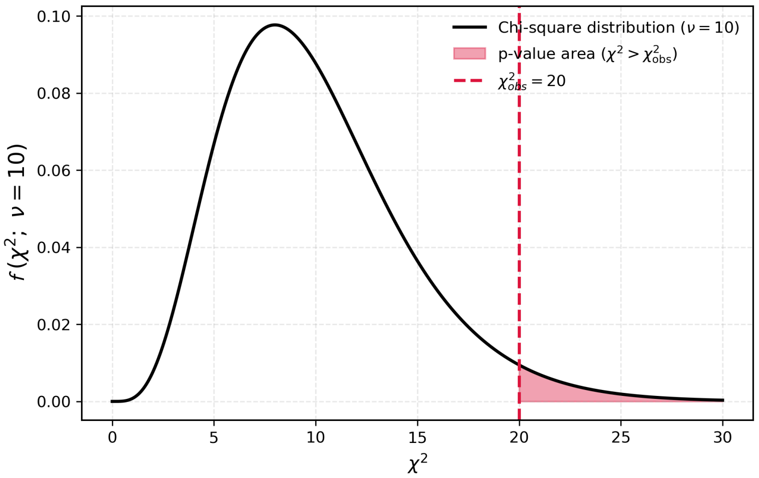
(b) Representation of the 1-sided P-value, computed as the integral of the right tail of the t distribution.



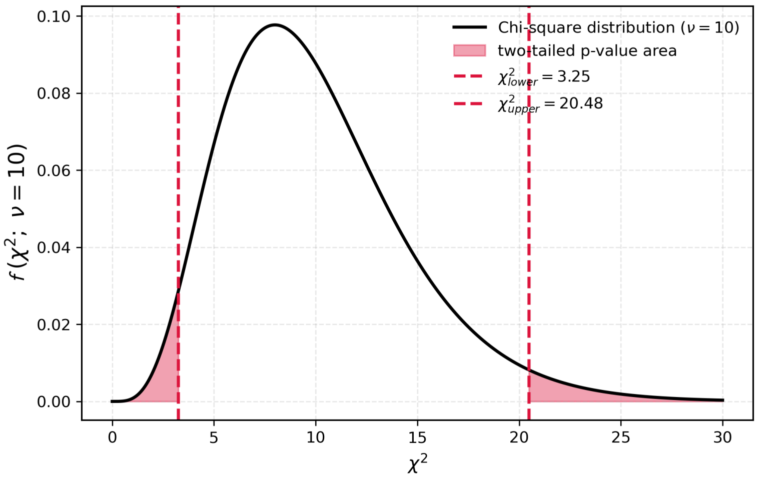
(c) Representation of the 2-sided P-value. Given the symmetry of the t-distribution, it can be obtained as double in size of the 1-sided integral.



(a) The Pearson's χ^2 distribution of the χ^2-statistic, given different values of the degrees of freedom ν.



(b) Representation of the 1-sided P-value, computed as the integral of the right tail of the t distribution.



(c) Representation of the 2-sided P-value. Given the symmetry of the t-distribution, it can be obtained as double in size of the 1-sided integral.

Chapter 5

Introduction to conditional probability

> *Probability statements are just summaries of repeated observations.*
>
> — W. V. Quine

The topic of conditional—sometimes referred as *bayesian*—probability has its roots in one most fundamental princples know by human nature. That is, the idea that we all have bias, and that purely objective knowledge is beyond our reach.

5.1 Motivation and philosophy

5.2 Dependent and independent events

5.3 Some examples of conditional probability

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Appendix A

Appendix 1

The integral

$$\int_a^b f(x)\, dx = \lim_{n\to\infty} \sum_{i=0}^{n} f(x_i)\, \Delta x \tag{A.1}$$

Equivalently

$$\int_a^b f(x)\, dx = \lim_{n\to\infty} \sum_{i=0}^{n} f(x_i)\, \Delta x \tag{A.2}$$

Appendix B

Appendix 2

Additional examples and computations may be placed here.

Appendix C

Appendix 3

Additional examples and computations may be placed here.

Bibliography

[1] John P. A. Ioannidis. “why most published research findings are false”. *PLoS Medicine*, 2(8):e124, 2005.

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

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

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

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

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

[7] Irving L. Finkel. “the ancient origins of dice". *Antiquity*, 81(314):176–187, 2007.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[26] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, 1933.

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

[28] Ronald A. Fisher. On a distribution yielding the error functions of several well known statistics. *Proceedings of the International Congress of Mathematicians*, 1924.

[29] Ronald A. Fisher. *Statistical Methods for Research Workers.* Oliver and Boyd, 1925.

[30] Ronald A. Fisher. *The Design of Experiments.* Oliver and Boyd, 1935.

[31] Karl Pearson. 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*, 50:157–175, 1900.

[32] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54:426–482, 1943.

[33] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[34] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[35] Howard Levene. Robust tests for equality of variances. *Contributions to Probability and Statistics*, pages 278–292, 1960.

[36] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

[37] Andrey Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 1933.

[38] Nikolai V. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.

[39] Samuel S. Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611, 1965.

[40] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, 1933.

[41] Hans Reichenbach. *Experience and Prediction.* University of Chicago Press, 1938.

[42] Karl Popper. *Logik der Forschung.* Springer, 1934.

[43] Erich L. Lehmann. *Testing Statistical Hypotheses*. Wiley, 1959.

[44] David R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006.

[45] Deborah G. Mayo. *Error and the Growth of Experimental Knowledge*. University of Chicago Press, 1996.

[46] Deborah G. Mayo. *Statistical Inference as Severe Testing*. Cambridge University Press, 2018.