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

Jesús Urtasun Elizari

January 2, 2026

Contents

Preface

The purpose of these notes

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

Both, predictive probability and descriptive statistics have deep historical roots, from ancient works on chance and divination to modern scientific topics oriented towards information theory, modelling and data analysis. As one could guess, rivers of ink have been written about such topics, and endless literature sources are available. However, after following many different courses at both bachelor and postgraduate levels, and teaching such topics myself during the last three years, I have found that most resources belong, almost certainly, to one of the next three classes. Either (*i*) deeply mathematical, and hence out of reach for most experimental or clinically oriented scientists, (*ii*) laboratory oriented, focusing on inference and experimental design, and hence missing most of the mathematical background, or (*iii*) with a direct focus towards programming and computation, relying on domain specific notebooks (Python, R, Matlab, SPSS, etc), and online resources with precompiled libraries for simulation, which again miss most of the mathematical and formal intuitions. Indeed, the misuse of statistics in experimental sciences is a critical topic in modern times, as mathematicians have extensively discussed during the last decades. The well-known article by John P. A. Ioannidis, *"Why most published research findings are false"* [17], 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"* [27].
- A more foundational textbook, with more advanced mathematical approach, can be found at DeGroot and Schervish's *"Probability and Statistics"* [9].
- For a philosophical and historical perspective on probability and statistics, please find Forster and Bandyopadhyay's handbook *"Philosophy of Statistics"* [1].
- A comprehensive introduction with focus on practical applications and modern data analysis tools is can be found at Diez, Barr & Mine *"OpenIntro Statistics"* [11].
- 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"* [23].

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 [12]. 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"* [7].

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 [5]. 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"*) [4], 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 [10]. 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"* [16], 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 [3, 15].

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. We shall cover some basic notations of Markovian probability in Chapter 7 [...]

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"*) [18], 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.

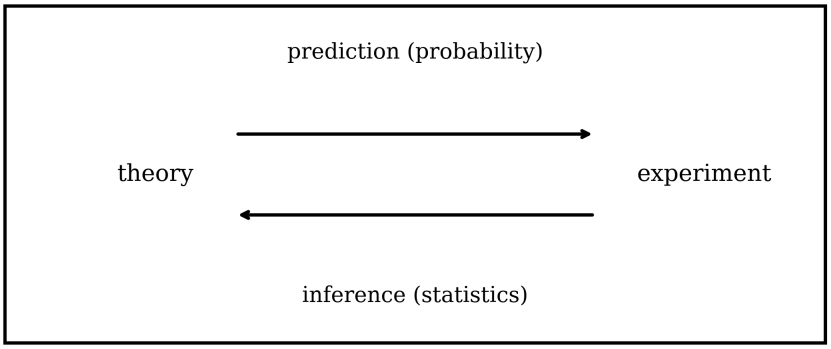


Figure 1.1: Representation of the predictive (from theory, or model, to experimental verification) and inferential (from data, measurement, observation to underlying truth) approaches to natural phenomena. As an example of the predictive branch of mathematics dealing with uncertainty we would find the 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.

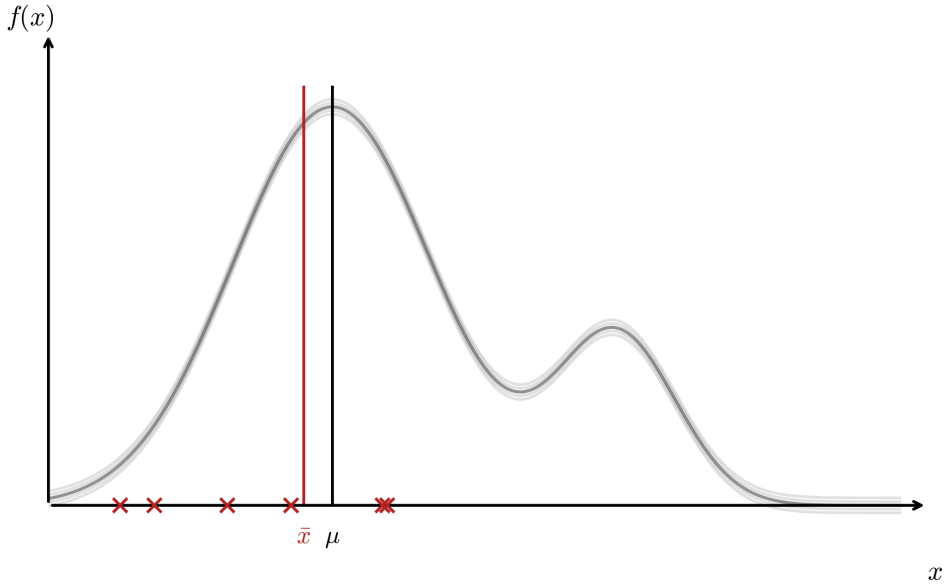


Figure 1.2: Representation of the *true* population mean μ, in black, and the observed *sample* mean $\bar{x}$. The true mean is and ideal and unaccessible quantity, while the sample mean can be computed as an estimator of the finite sample.

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

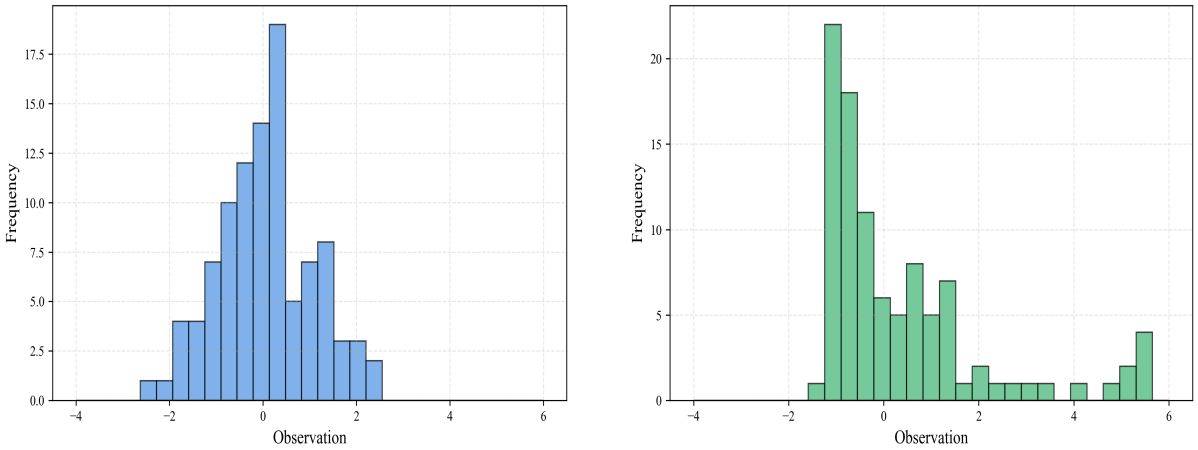
$$Q_p = \inf\{x : F(x) \geq p\}. \tag{1.8}$$

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 [...].

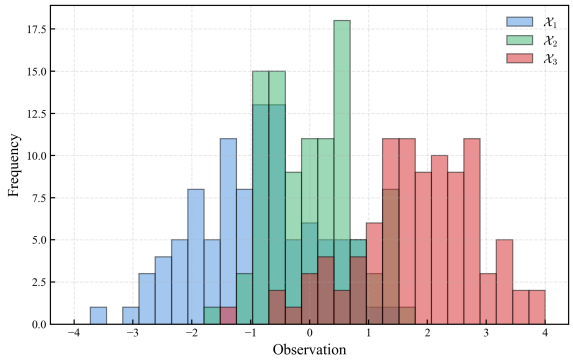
3. Exercise [...].

Solutions

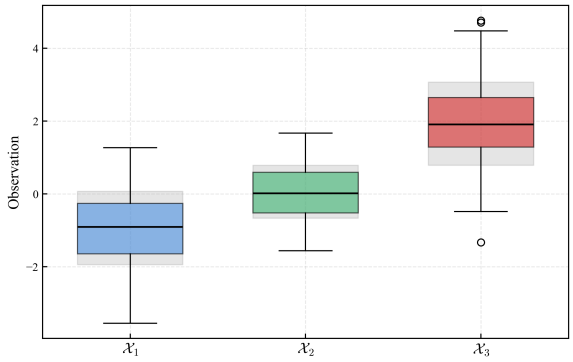
1. Solution [...].

2. Solution [...].

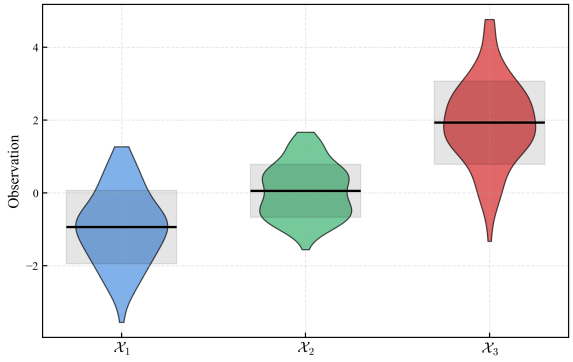
3. Solution [...].



(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 on games of chance, and in particular the "problem of the division of stakes," laid the groundwork for the systematic analysis of uncertain events. Years later, Jacob Bernoulli's *Ars Conjectandi* established the first classical definition of probability, providing the study of random events with mathematical clarity. Refinements by De Moivre and Laplace transformed it into a powerful analytical theory, while its true axiomatic structure only crystallised in the twentieth century with Kolmogorov's *Grundbegriffe der Wahrscheinlichkeitsrechnung* in 1933 [18].

The mathematical formalization of decision-making is actually quite a recent development. It is usually attributed to British mathematician Frank P. Ramsey (1903–1930), who in his 1926 paper *Truth and Probability* [25] introduced a formal, subjective interpretation of probability, laying the groundwork for what later became expected utility theory in decision-making under uncertainty. In short, Ramsey formalized how rational agents should assign probabilities and make decisions based on personal beliefs and preferences. All starting from the apparantley-simple question *'how should we make decisions in the face of uncertainty?'*.
to the twentieth-century developments of Pearson, Fisher, and Neyman addressed the latter rather than establishing the former distinction.

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

2.1 What is probability?

At its heart, probability is nothing more - and nothing less - a branch of mathematics developed to describe random events, also referred to as *stochastic*. Indeed, the word "stochastic" comes from the Greek word στοχαστικός, which literally means "to guess" or "to aim." The way we describe such events, characterized by the uncertainty of their outcome, is by defining a

quantity we will call $\mathbb{P}$, of probability. That quantity $\mathbb{P}$ will denote a number between 0 and 1, which reflects the degree of uncertainty, or *surprise*, with which the random event produces a specific outcome. For an event A, such as observing a heads when tossing a coin, or a given face when rolling dice, the numerical convention is written as follows,

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

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

2.2 Discrete events

By *discrete* we mean that the number of possible outcomes is a finite or countable number [...]. In such cases, distributions will exactly represent a probability, and they are referred to as *mass* distributions [...]

2.2.1 Bernoulli trials

2.2.2 Binomial distribution

2.2.3 Poisson distribution

2.2.4 Discrete uniform distribution

2.3 Continous events

By *continous* we mean that the number of possible outcomes is an infinite, uncountable number, in a continuous range [...]. In such cases, we will need to build a new mathematical object, referred as a probability *density* [...]

2.3.1 Gaussian distribution

2.3.2 Exponential distribution

2.3.3 Continous uniform distribution

2.4 Expected values

Expectation. The *expected value* (or mean) of a random variable formalizes the idea of a probability–weighted average of all possible outcomes. For a discrete random variable X with probability mass function p_X, the expectation is obtained by summing each possible value weighted by its probability. For a continuous random variable with density f_X, the sum is replaced by an integral. In both cases, expectation is defined whenever the corresponding series or integral converges absolutely.

Variance. The *variance* measures dispersion around the mean. It is defined as the expected squared deviation from the mean,

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

Expanding the square yields the useful identity

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

which simplifies most computations and highlights that variance depends on the second moment of the distribution.

Discrete Distributions

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

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1-p) = p,$$

$$\operatorname{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p).$$

Binomial distribution. Let $X \sim \mathrm{Bin}(n,p)$. Writing $X = \sum_{i=1}^{n} X_i$, where the X_i are i.i.d. Bernoulli(p),

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np,$$

$$\operatorname{Var}(X) = \sum_{i=1}^{n} \operatorname{Var}(X_i) = np(1-p).$$

Discrete uniform distribution. Let X be uniformly distributed on $\{1, \dots, n\}$.

$$\mathbb{E}[X] = \frac{1}{n}\sum_{k=1}^{n} k = \frac{n+1}{2},$$

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{k=1}^{n} k^2 - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}.$$

Poisson distribution. Let $X \sim \mathrm{Pois}(\lambda)$. Using the series definition of the exponential function,

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \lambda, \qquad \operatorname{Var}(X) = \lambda.$$

Continuous Distributions

Gaussian distribution. Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Symmetry of the density about μ implies $\mathbb{E}[X] = \mu$. A direct computation of $\mathbb{E}[(X-\mu)^2]$ using Gaussian integrals yields

$$\operatorname{Var}(X) = \sigma^2.$$

Exponential distribution. Let $X \sim \mathrm{Exp}(\lambda)$ with density $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

$$\mathbb{E}[X] = \int_0^{\infty} x \lambda e^{-\lambda x}\, dx = \frac{1}{\lambda},$$

$$\operatorname{Var}(X) = \int_0^{\infty} x^2 \lambda e^{-\lambda x}\, dx - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

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

$$\mathbb{E}[X] = \frac{1}{b-a}\int_a^b x\,dx = \frac{a+b}{2},$$

$$\text{Var}(X) = \frac{1}{b-a}\int_a^b (x-\mu)^2\,dx = \frac{(b-a)^2}{12}.$$

Conceptual Summary

Expectation is a *first-moment* quantity: it captures location but ignores spread. Its defining feature is linearity,

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

which holds without independence and underlies most statistical estimators.

Variance is a *second-moment* quantity, measuring dispersion relative to the mean. Unlike expectation, it is not linear, but it is additive for independent random variables. This property explains the appearance of variances in limit theorems and error propagation.

Together, mean and variance form the basic quantitative summary of a distribution. Classical inference procedures—such as confidence intervals, t-tests, and Wald statistics—are built on estimators of these moments and on their sampling distributions.

Chapter 3

Prediction, inference, sampling distributions

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

— Florence Nightingale

Let's revisit again the difference between prediction and inference, as is through estimation that both, probability and inference become part of a two-folded problem.

3.1 Prediction vs inference

3.2 The Law of Large Numbers

3.3 The Central Limit Theorem

3.4 Bias, variance and Mean Squared Error

3.5 Confidence intervals and critical regions

3.6 Application to Generalized Linear Models

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Chapter 4

Introduction to hypothesis testing

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

— Ronald A. Fisher

The term *hypothesis testing* lies on top of the two pillars we have mentioned in previous chapters. On the one hand, we will use the basic statistical analysis tools we described in Chapter 1, such as sampling, estimators and general of data visualization. At the same time, we will rely on probability theory to predict expected values about the true popuplation parameters, assuming certain distributions, etc, following what we discussed in Chapter 2. Most of our examples will assume that our data is simple, smooth, and Gaussian distributed—what people normality refer to as *parametric*—given the Law of Large Numbers and the Central Limit Theorem, and building confidence intervals and critical regions to ensure our estimators—sample mean and variance—reliably represent the population under study, quantifying their central tendency and variation. All these have been discussed in Chapter 3.

With all these properly covered, we can now start formulating and testing hypotheses. With a predictive mathematical theory, such as probability and combinatorics, we can compute expected values for the true population mean or variance of a given population. With statistical analysis we can build estimators that quantify central tendency and variation, and visualize distribution and outlier behaviors. Finally, we can rely on the results given by the LLN and CLT, and confidence intervals to ensure that bout our expectations and data is smooth and simple to address. With all these, we will be able to to define a new type of *informative quantity* normally referred to as *statistic*, or *statistic test*, that quantifies how much our observed data approaches—or differs from–the expected or hypothesized value. Finally, following the so-called modern, or Pearson-Neyman approach to hypotheis testing, we will learn how to quantify significance through the computaiton of the Pearson value—or P-value, for short.

Before rushing to mathematical or technical definitions, we would like to pencil a brief historical note, that we hope will shed some light on this topic, sometimes rather obscure. The very idea of hypothesis, and verification against experimental observations, is indeed old, and can be traced back to [...]. But quantities such as the arithmetic mean or other estimators we discussed in previous chapters were not properly defined or become a standard in scientific research until quite later [...]. Similarly, the very foundations of probability theory, including the idea of random variable, unitarity, or distributions—Uniform, Binomial, Gaussian—, were not properly until the work of Kolmogorov in 1933 [...]. The problem of mathematically quantify how much an expectation matches or approaches empirical data, is indeed even younger. It is only since the early 1920s with the works of Karl Pearson and Ronald Fisher, among others, that the very idea of statistic tests were developed, and indeed quite different that the modern approach we are used to nowadays. It was Fisher in the (1922, 1925) connected least squares,

likelihood, and sampling distributions, establishing the foundations of modern inference. The original development of either the χ^2, the t-test, or the Fisher F-test, *was not linked to the acceptance / rejectance of hypotheis based on P-values.* The whole philosophy of relying on P-values to accept / reject null hypotheses, and the very formulation of null / alternative hypotheis on the first place, traces back to Egon Pearson and Neyman (1937), as part of their work on formalizing the idea of confidence intervals as frequentist procedures. The meaning of all these quantifies and their interpretation remains nowadays an open philosophical debate, hence we will define them carefully and walk throug each of them with clear examples. We will revisit the interpretation and historical discussion further in this chapter, and hope this brief disclamer will help to start dismantling now preconceived notions, and approach the topic carefully.

4.1 Prediction vs inference revisted

When formulating hypothesis about natural phenomena, we shall remember once again the difference betwen population and smapling. On the one hand, we have an idealized, unaccessible population with unaccessible true mean, variance, etc. The way I can compute a mathematical prediction for a random variable, whatever it represents, is through the calculation of an expected value—also referred to *momentum*—given a random variable and its distribution.

Given a random variable, we can compute expected values, or momenta of the distriutions:

Population mean for a discrete random variable x_i

$$\mu = \mathbb{E}[x] = \frac{1}{N}\sum_{i=1}^{N} x_i \, \mathbb{P}(x_i) \tag{4.1}$$

Population variance for a discrete random variable x_i

$$\sigma^2 = \mathbb{E}[x-\mu] = \frac{1}{N}\sum_{i=1}^{N} (x_i-\mu)^2 \, \mathbb{P}(x_i) \tag{4.2}$$

Population mean for a continous random variable x

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

Population variance for a continous random variable x

$$\mu = \mathbb{E}[x-\mu] = \int_{i=-\infty}^{\infty} (x_i-\mu) * \ f(x) \, dx \tag{4.4}$$

if x is a continous random variable.

Hence hypotheses will *always be formulated in terms of a mathematical predictions about the population parameters.* If I believe in Newtonian mechanics, a hypothesis could be to write down Newont's second law and use it to predict where and when a stone would fall when dropped from a certain height—its position and time. If my hypothesis is that a gene has a cerain impact in a known disease, or in response to stress, —its expression level, or counts. Or, if I am studying the relation between smokers in the UK and their probability to develop lung cancer [...]. In any of these cases, upon hypothesis. I would need samples, or groups, of measurements, normally referred to simply as *data.*

Observed sample mean for a sample $\chi = \{x_1, x_2, ..., x_n\}$

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

Observed sample variance for a sample $\chi = \{x_1, x_2, ..., x_n\}$

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

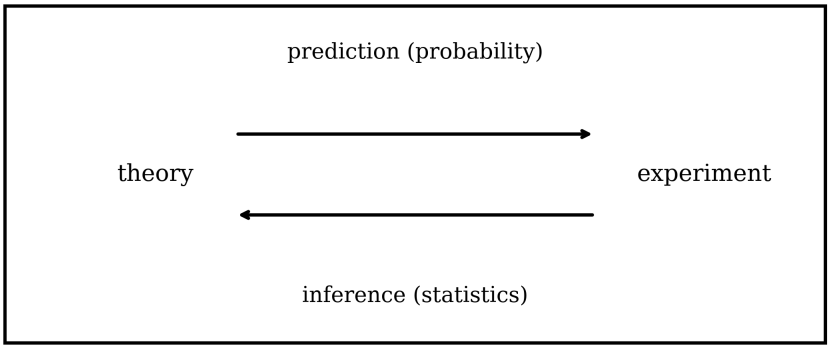


Figure 4.1: Representation of the predictive (from theory, or model, to experimental verification) and inferential (from data, measurement, observation to underlying truth) approaches to natural phenomena. As an example of the predictive branch of mathematics dealing with uncertainty we would find the 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 hypothesis, predictions, experiments and data, there is plenty of approaches and formulations, as many as instruments, scales and fields of study. These do change form one field to another, and they do change in time. Our very idea of hypothsis, prediction, measurement, and law [...]. Nowadays, when people refer to *hypothesis testing* they mean very specific approach, almost and algorithmc-wise set of rules, that is applied in general to inference and data science problems. We will define such approach as the "modern", or "general" approach to hypothesis testing, that asumes some basic notions of probability theory, distributions and randomness, with some of statistics, estimators and sample description [...]. The whole idea of statistic test, P-value and significance, that we will discuss now, ranges indeed from quite recent times, back to Pearson, Fisher, and Neyman in the early 1900s.

- Formulate hypothesis. Normally referred to as *null* hypothesis H_0, as the expectation that our prediction or expectation will follow, and *alternative* hypothesis H_1, representing the case of finding a surpsing observation, that deviates from H_0. These hypothesis will *always* be made about the *true population parameters*, and commonly formulated as the computation of an expected value, that we discussed in Chapter 2.

- Experiment, measurement, observation. Any process, regardless of instrumentation and object of study, that involves a measurement, an observation, or data collection of any kind from one or more samples.

- Compute statistic, or statistic test. Out of our random data we can caompute any *informative quantity*, which can be an estimator like the sample mean, the variance, etc, or a more abstract quantity that represents how close are the these mean and variance from their expected values, given H_0

- Copute P-value: the probability that, given a certain assumtion for our true population parameters and our random data, we obtained a value at least as extreme as the one we got for our statistic.
- interpretation of the result, normally accept / reject the null hypothesis based on the P-value and some significance threshold.

A couple of notes about this general roadmap. A statistic can be just an estimator, like the sample mean [...]. Fisher's definition of P-value as extream [...]. The approach is a mixed of Fisher and Pearson-Neyman [...].

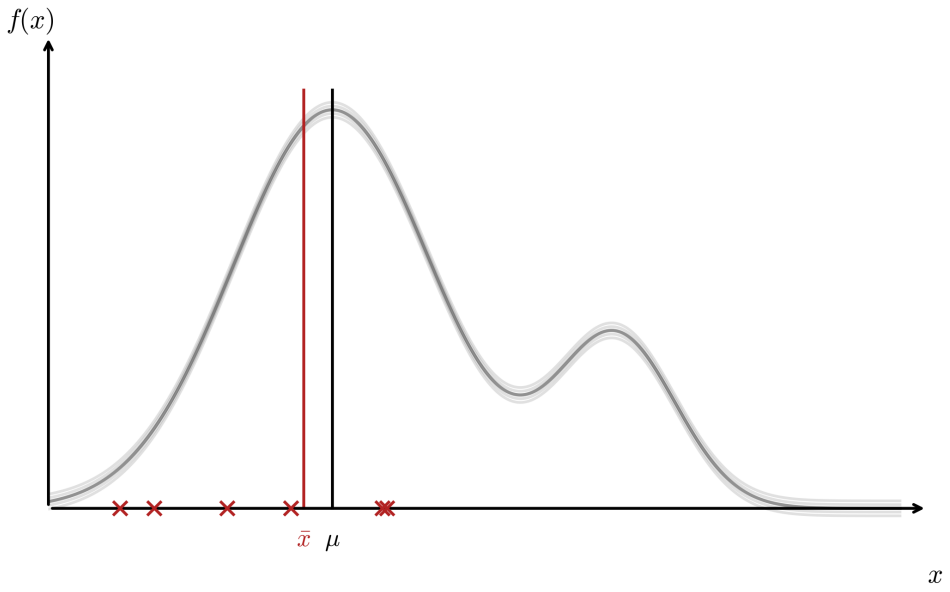


Figure 4.2: Representation of the *true* population mean μ, in black, and the observed *sample* mean $\bar{x}$. The true mean is and ideal and unaccessible quantity, while the sample mean can be computed as an estimator of the finite sample.

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 statistic test we will discuss. It was developed in 1908 by William S. Gosset, a statistician working at the Guinness factory in Dublin, 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 to the Guinness company he has not allowed to publicly share his work and hence he submitted it to the *Biometrika* statistics journal under the pseudonym *Student*. This is why it remains nowadays known as the *Student's t-test*.

The test begins by formulating some null hypothesis about the true population mean, *prior to any sampling or data collection*. Normally, the null hypothesis is simply written as *the true population mean is expected to take the value* μ. It is important here to stop and think carefully about what is the physical quantity that we are actually going to measure. Remember that such quantity, unaccessible in theory, 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 observations, or measurements, and group them in a given sample $\chi = \{x_1, x_2, \ldots, x_n\}$. Out of them, we can compute the sample mean $\bar{x}$, as an estimator of

the true population mea 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, the sample mean $\bar{x}$ and the sample standard deviation s, as our estimators) we can compute the *t-statistic*, or *t-statistic test*, defined as

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

Let's look at this quantity for a second. We will notice that as the sample mean $\bar{x}$ approaches the expected value μ, the t-variable tends to zero. It was designed for this precise purpose, to quantify how different—or similar—our data is from the hipothesized value, and in the ideal case $\bar{x} \to \mu$, then $t \to 0$.

It is important to note here that t 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 of $\bar{x}$ and s, hence producing a different t. This means that t *is a real-valued random variable itself*, and it will follow *some* distribution. The mathematical definition of such distribution—or density—of the t-variable, as introduced in Gosset's work [...], is called the *Student's t-distribution*,

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

where the parameter ν is referred to as the *degrees of freedom*, and it is simply related to the length of the sample $\nu = n - 1$. You may see that some textbooks and literature sources write it as t_ν, which is perfectly fine and common standard, but we prefere to use here the letter t just for the statistic, and $f(t;\nu)$ for the distribution of t given ν degrees of freedom, rather than referring to both elements with the same symbol. It can be demonstrated that as the degrees of freedom increase, the t-distribution tends asymptotically to a Gaussian distribution, normally written as $f(t;\nu) \to \mathcal{N}(0,1)$ for $\nu \to \infty$. The explicit demonstration is out of the scope of this course, but it can be found at [...].

We arrive now to the final step; the computation of the P-value. Back to the well known Fisher's definition of P-value, *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 see that once known the distribution of the t-statistic, we just need to compute the cumulative probability—that is, the integral—of such distribution.

If we ask for the probability of obtaining a value *strictly greater or strictly smaller than the one we observed for our statistic*, $\mathbb{P}(t \geq t_{\text{obs}})$, we just need to compute the integral of one of the distribution tales, and it is referred to as a *one-sided*—or one-*tailed*—P-value. On the other hand, if we ask for the probability of obtaining a value *more extreme, regardless of the direction, than the one we observed*, $\mathbb{P}(|t| \geq |t_{\text{obs}}|)$, that would require the integral of both tails, and it is referred to as a *two-sided*—or two-*tailed*—P-value. Given the symmetry of the t-distribution, this reduces to double the size of 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 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. For a sample of size $n = 10$, 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 $p \approx 0.54$.

4.3.2 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 [...].

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*,

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

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 case the two-sample t-test was an extension of the one-sample case, both wollowing Gosset's formulation of the statistic and corresponding distribution, Fisher's Analysis of Variance, or ANOVA test, is a specific case of the general F test we just discussed. It was developed as part of his investigation of the sampling distributions of quadratic forms under normality, aiming to check whether two sources of variability could plausibly be attributed to the same underlying variance, without an explicit parameter-first formulation of hypotheses. Rather than comparing two variances in isolation, Fisher decomposed total variability into components attributable to multiple factors and experimental design.

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

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

$$F = \frac{s^2_{\text{between}}}{s^2_{\text{withinn}}} \ . \tag{4.12}$$

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.

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. Pearson's original formulation was oriented towards evaluating distriutions and quantify similarly, rather than formal hypothesis testing: Given a list of observations $\{O_1, O_2, \ldots, O_n\}$ and a series of expectations $\{E_1, E_2, \ldots, E_n\}$, the χ^2 statistic

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \tag{4.13}$$

was conceived as a numerical measure of discrepancy between observed and expected frequencies under a proposed model. Large values of χ^2 indicated poor agreement, while a χ^2 close to zero indicates point-wise similarly. [1]

Same as in the two previous examples, the χ^2-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 O_i, hence producing a different χ^2. This means that χ^2 *is a real-valued random variable itself*, and it will follow *some* distribution. The mathematical definition of such distribution—or density—of the χ^2-variable, as introduced by Pearson in his work [...], is called the Pearson's χ^2*-distribution*,

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

where χ^2 is a strictly positive quantity, and the degrees of freedom ν depend on the number of constraints imposed by the data structure and on the number of parameters estimated under the null hypothesis, unlike the t-test where the degrees of freedom are always $n-1$ due to the estimation of a single mean, now they are defined as.

- When used to evaluate the goodness-of-fit test (multinomial, with k categories) $\nu = n-1-p$, where p is the number of parameters estimated from the data. In particular, $\nu = n-1$ if no parameters are estimated, and $\nu = n-2$ if one parameter is estimated (e.g. a mean).
- When used to test of independence in a contingency table ($r \times c$). $\nu = (r-1)(c-1)$. In all cases, the degrees of freedom reflect the effective number of independent components remaining after accounting for normalization constraints and parameter estimation.

In the modern framework, the chi-square test is formulated with an explicit null hypothesis and a P-value. One specifies

$$H_0: \text{ the observed frequencies follow the model } F_0$$

(or independence in a contingency table), derives the asymptotic distribution

[1] Pearson's original work does not distinguish null and alternative hypotheses in the modern sense, nor does it invoke Type I and Type II errors; these concepts were introduced later by Neyman and Pearson in the late 1920s and early 1930s.

$$\chi^2 \sim \chi^2_\nu \quad \text{under } H_0,$$

and evaluates significance via the tail probability of the observed statistic. This reformulation transformed Pearson's discrepancy measure into a decision-theoretic test with controlled error rates, fully aligned with modern parametric inference.

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

$$P_{\text{one-sided}} = \mathbb{P}(\chi^2 \geq \chi^2_{\text{obs}}) = \int_{\chi^2_{\text{obs}}}^{\infty} f(\chi^2; \nu)\, d\chi^2,$$

$$P_{\text{two-sided}} = 2\min\left\{ \int_0^{\chi^2_{\text{obs}}} f(\chi^2; \nu)\, d\chi^2,\ \int_{x_{\text{obs}}}^{\infty} f(\chi^2; \nu)\, d\chi^2 \right\}.$$

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 [...].

A deeper theoretical connection appears through likelihood-based testing. In many settings, Pearson's chi-square statistic is asymptotically equivalent to the likelihood-ratio statistic

$$-2\log\Lambda,$$

a result formalized by Wilks in the 1930s. Both statistics converge in distribution to a chi-square law under the null hypothesis, reflecting the quadratic approximation of the log-likelihood near the true parameter. From this perspective, the classical chi-square test can be viewed as an early large-sample approximation to likelihood-based inference, with likelihood-ratio, Wald, and score tests providing a unified parametric framework that generalizes Pearson's original idea.

4.3.6 The Wald test: asymptotic behavior

The Wald test was developed in the 1930s by Abraham Wald in the context of general parametric inference and decision theory. 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. It thus unifies earlier procedures within a common asymptotic theory.

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

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

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. 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, independently introduced by Mann and Whitney in 1947, provides a non-parametric alternative to the two-sample t-test for independent samples. 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. 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.

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. 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. 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 [13, 14]. 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 [22]. 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 [2] and developed by Laplace and later subjectivists such as de Finetti [8], 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.

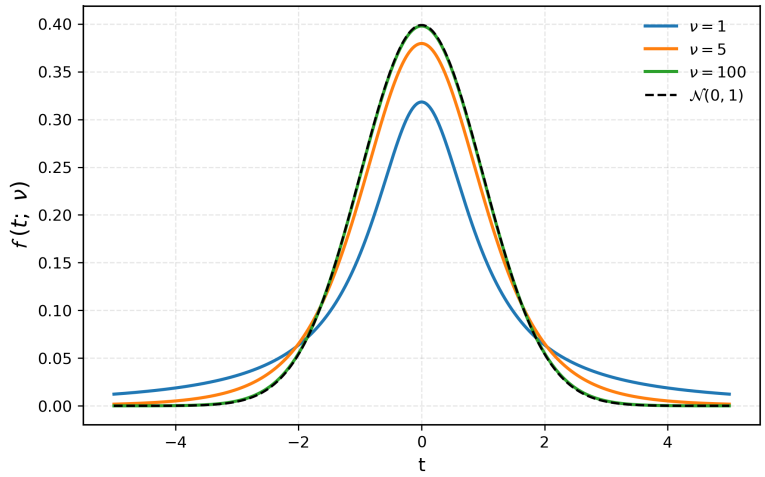
Hans Reichenbach provided the clearest philosophical articulation of the frequentist stance underlying Neyman–Pearson theory [26]. He distinguished *prediction*—statements about long-run frequencies—from *inference*—claims about truth or belief. Statistical tests, on this view, justify actions and predictions 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.

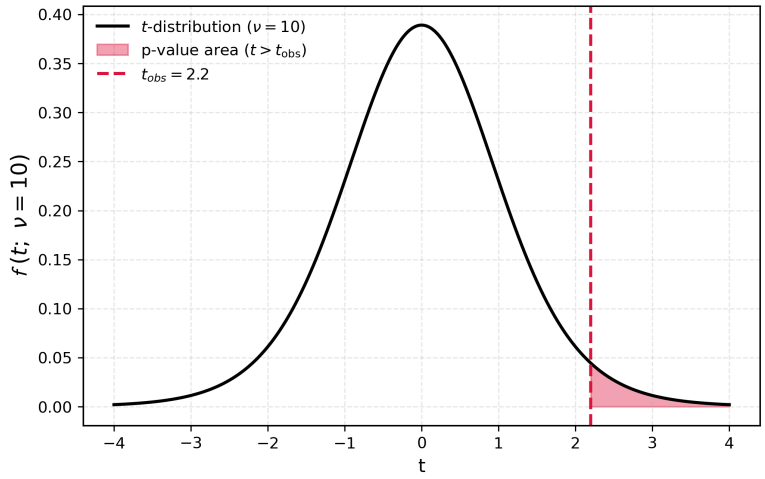
Karl Popper rejected probabilistic confirmation altogether, arguing that science advances through bold conjectures and severe attempts at falsification [24]. Statistical tests, in his view, contribute by formulating risky predictions whose failure can refute theories, not by accumulating evidence or 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 [19], 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 [6]. Deborah Mayo further developed an error-statistical philosophy in which evidential interpretation is grounded in the severity with which hypotheses are tested [20, 21]. 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.

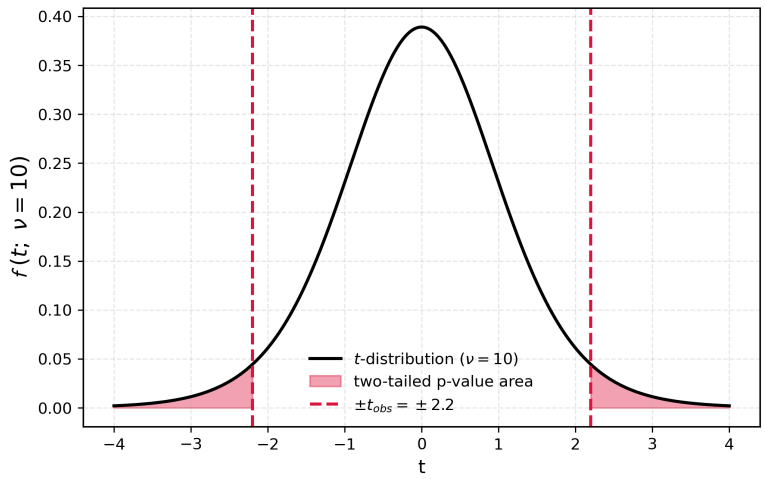
The coexistence of Fisherian evidence, Neyman–Pearson decision rules, Bayesian belief updating, Popperian falsification, and Reichenbach's predictive frequentism reflects not confusion but plurality. Each framework answers a different question—about evidence, action, belief, or prediction—and Type I 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.



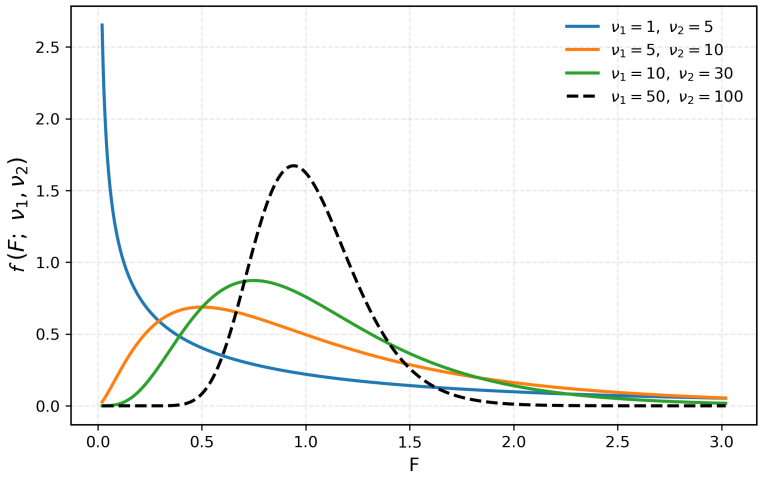
(a) The Student's t distribution of the t-statistic, given different values of the degrees of freedom ν.



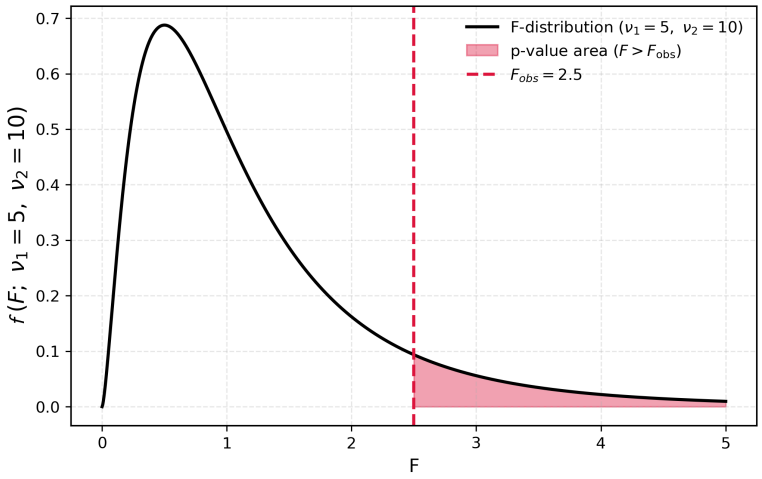
(b) Representation of the 1-sided P-value, computed as the integral of the right tail of the t distribution.



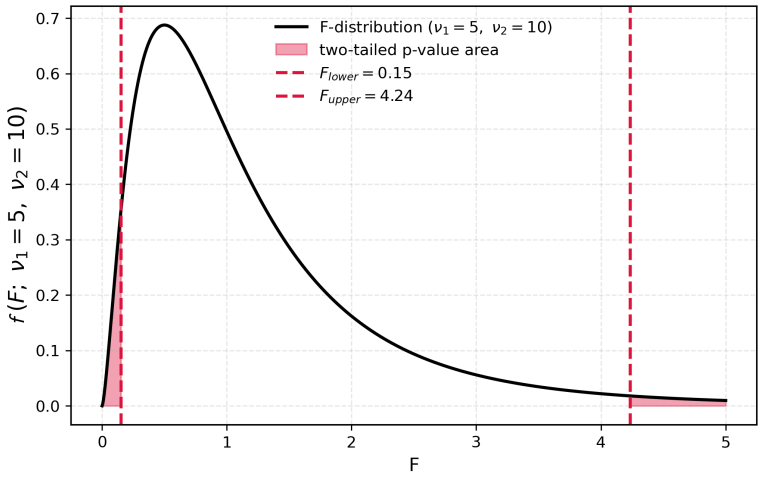
(c) Representation of the 2-sided P-value. Given the symmetry of the t-distribution, it can be obtained as double in size of the 1-sided integral.



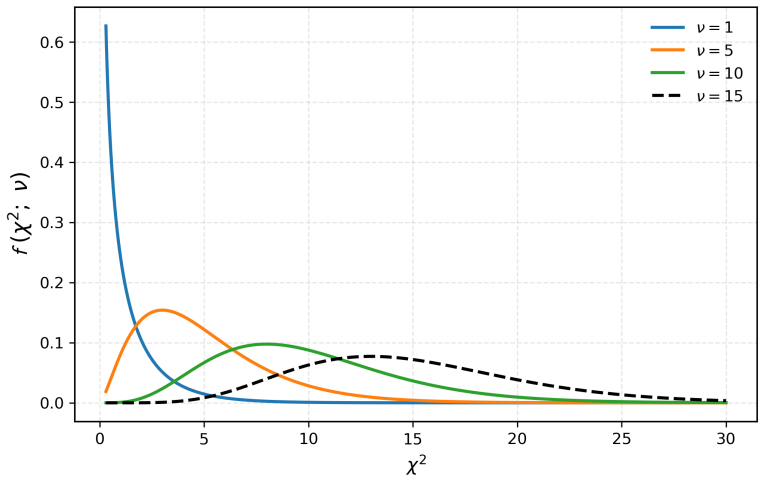
(a) The Fisher's F distribution of the F-statistic, given different values of the degrees of freedom ν.



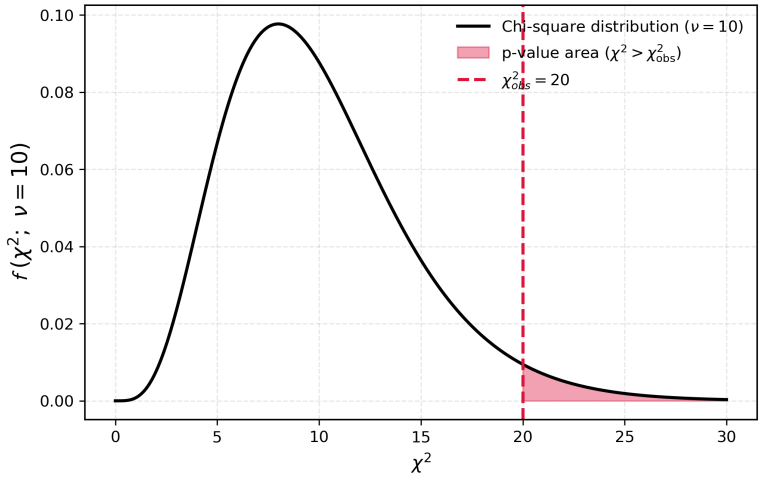
(b) Representation of the 1-sided P-value, computed as the integral of the right tail of the t distribution.



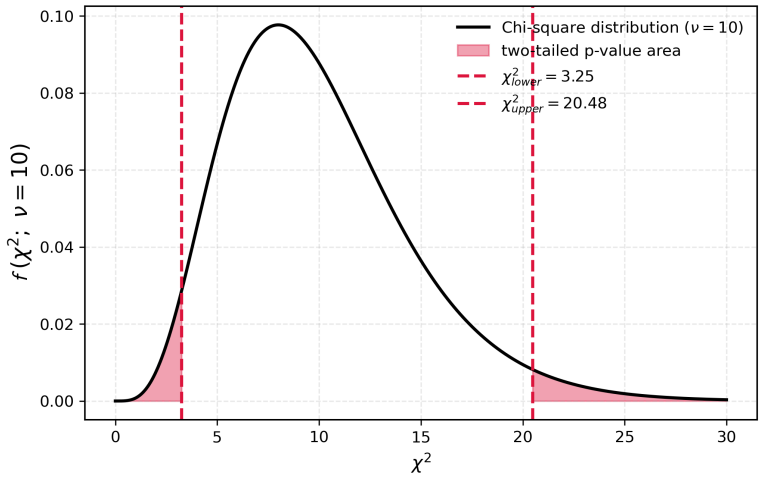
(c) Representation of the 2-sided P-value. Given the symmetry of the t-distribution, it can be obtained as double in size of the 1-sided integral.



(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

Modelling, dependency and correlation

The theory of probabilities is at bottom nothing but common sense reduced to calculation.
— Pierre-Simon Laplace

5.1 Introduction and Philosophy

Matrix-based linear modelling was systematized in the mid-20th century, notably in the work of C. R. Rao (1945, *Bulletin of the Calcutta Mathematical Society*), who developed the Cramér–Rao bound and unified estimation in linear models.

5.2 Estimation and Inference

Model estimation chooses parameter values that best describe the data; inference quantifies uncertainty around these estimates.

Mathematical Formulation

The ordinary least squares estimator is

$$\widehat{\beta} = (X^\top X)^{-1} X^\top \mathbf{Y},$$

with residuals $\widehat{\varepsilon} = \mathbf{Y} - X\widehat{\beta}$. Under the normal-error model,

$$\widehat{\beta} \sim \mathcal{N}\big(\beta, \sigma^2 (X^\top X)^{-1}\big).$$

Numerical Example

For

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \qquad Y = \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix},$$

one obtains

$$\widehat{\beta} = \begin{pmatrix} 0.333 \\ 1.5 \end{pmatrix},$$

so $\widehat{Y} = 0.333 + 1.5X$.

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Chapter 6

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.

6.1 Motivation and philosophy

6.2 Dependent and independent events

6.3 Some examples of conditional probability

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Chapter 7

Stochasticity and Markov Processes

The development of mathematics is a continuous process of abstraction.

— Emmy Noether

To end this manuscript, we—plural de cortesía—would like to introduce a topic of growing interest in the present years, because of its deep implication—among many others, much less known—than LLMs or AI-related applications.

7.1 Motivation and philosophy

7.2 Mathematical definition

7.3 Some examples of conditional probility

7.4 Stochasticity and Markov processes

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[18] Andrey Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933.

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

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

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

[22] 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.

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

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

[25] 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.

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

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