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

Jesús Urtasun Elizari

December 12, 2025

Contents

Preface

This is a minimal example of a probability and statistics book. The purpose of this preface is simply to allow the document to compile cleanly.

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 [11]. 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 [6]. 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"*) [5], posthumously published in 1663, is one of the earliest known texts to explore probability through the analysis of gambling problems. However, Cardano's reasoning, while insightful, lacked the symbolic clarity and mathematical rigour of modern probability theory. Readers consulting the original manuscript will notice an ambiguous and sometimes inconsistent symbolic system, quite unlike the formal structures we use nowadays.

The formalization of probability as a mathematical discipline did not occur until the 17th century, most notably through the seminal correspondence between Blaise Pascal and Pierre de Fermat. Their work, motivated by problems such as finding a fair division of stakes in interrupted games of chance, introduced foundational concepts such as combinatorics, expected value, and variance [9]. 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"* [14], 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, 12].

During the 19th century, probability theory began to intertwine more deeply with statistics and the emerging mathematical analysis of physical phenomena. Florence Nightingale, best known for her pioneering role in modern nursing, also made significant contributions to statistical methodology and graphical representation of data. Her use of polar area diagrams and 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 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, economics, and modern machine learning.

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

Chapter 1

Descriptive statistics

Statistics is the grammar of science.

— Karl Pearson

In this chapter, 1 we will discuss the nature of prediction and inference, population and sampling, and estatistical estimators. (...) A large part of history of science could be summarized as an effort to translate observations of reality into precise, mathematical understanding. A record of the continuous human striving for a formulation and description of the real world in mathematical terms. To define mathematically the phenomena we find in the natural world, it is necessary to develop tools that relate the one or more relevant quantities - sometimes called *variables* - and how they relate or change depending on one another. The purpose of modelling might be, for instance, to determine the distance from the earth to the sun, to estimate the number of stars in the observable universe, relating the boiling point of water to the external pressure, or the number of lung cancer patients to pollution levels around smoking areas.

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

Hence, a philosophical position often adopted is that statistics is essentially the study of uncertainty, and that the statistician's role is to assist workers in other fields who encounter uncertainty in their work. In practice, there is a restriction in that statistics is ordinarily associated with data; and it is the link between the uncertainty, or variability, in the data and that in the topic itself that has occupied statisticians. Statistics does not have a monopoly of studies of uncertainty. Probability discusses how randomness in one part of a system affects other parts.

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

The mathematical formalization of decision-making is actually quite a recent development. It is usually attributed to British mathematician Frank P. Ramsey (1903–1930), who in his 1926 paper *Truth and Probability* [19] introduced a formal, subjective interpretation of probability, laying the groundwork for what later became expected utility theory in decision-making under

uncertainty. In short, Ramsey formalized how rational agents should assign probabilities and make decisions based on personal beliefs and preferences. All starting from the apparantley-simple question *'how should we make decisions in the face of uncertainty?'*.

1.1 Sampling and data types

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

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

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

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

where the i-subscripts remind us that the sample consists of selected observations from the population, not necessarily consecutive or all of them. The population represents the ideal object of inference, while the sample is the concrete, finite evidence available to us. This distinction is far from trivial; a poorly chosen sample often misrepresents the population and may induce bias, whereas a carefully constructed one mirrors its essential features, and can be used to describe the underlying nature.

Equally important is the recognition that not all data are of the same kind. A common distinction is to consider *categorical* and *numerical* data. Categorical - or *qualitative* - data describes qualities or labels such as the eye colour of students in a classroom (blue, brown, green), the brand of a purchased smartphone, etc. Sometimes they are further divided into *nominal* categories, with no natural order, like the eye colour or the smartphone brand, and *ordinal* categories with a meaningful order. Examples of these would be the finishing places in a race (first, second, third), survey responses ranging from *strongly disagree* to *strongly agree*, etc.

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

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

As a summary, sampling and proper description of data establish the ground upon which statistics is built. Before calculating, summarizing, or diving into inference, one must ensure that the information collected is both representative and properly understood. Without these

foundations, descriptive measures risk floating unmoored, detached from the reality they claim to represent. Accurate sampling and rigorous description will lead to a faithful representation of the phenomena under study and their relationships, detecting anomalies, and even building accurate predictions.

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

1.2 Central tendency and variation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note that here k is just an integer that helps locate the middle position of an ordered data set of size n. If the sample size n is even, we write $n = 2k$, while for n odd, we write $n = 2k+1$. In the case of an odd-sized sample, the median is just the middle-point, while for an even size, it is computed as the average of the two middle points. The mathematical definition (1.6) may seem a bit unnatural at first, so let's navigate it with a couple of examples. Consider the sample $\mathbf{x} = \{1, 2, 3, 5, 3, 2, 7\}$. First, we order the data:

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

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

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

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

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

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

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

Unlike the mean, the median is robust to outliers and skewed data, capturing the central position of the dataset even with repeated values. For instance, the data represented in LHS of Figure 1.1 will be accurately described by computing the mean, given its symmetric behaviour, while the one in the RHS will be better addressed with a median, accounting for the skewness and the presence of outliers.

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

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

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

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

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

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

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

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

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

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

As a warm up exercise, try to compute the variance for a second sample, let's say $\mathbf{x} = \{4, 5, 6\}$. By substituting in the general expression (1.11) you should get the result

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

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

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

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

and for the entire population,

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

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

Finally, *quantiles* divide the ordered data into equal proportions. The p-th quantile Q_p is the value below which a fraction p of the data lies. Special cases include the *first quartile* (Q_1, 25th percentile), the *median* (Q_2, 50th percentile), and the *third quartile* (Q_3, 75th percentile). Formally, for a continuous cumulative distribution function (CDF) F, the p-th quantile satisfies

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

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

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

1.3 Data visualization

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

The *histogram* is found among the oldest and most fundamental visualization tools. The concept of dividing data into intervals to visualize frequency dates back to Karl Pearson in the late 19th century, who formalized it as a graphical representation of probability distributions [18]. A histogram divides the range of a dataset into consecutive intervals, or *bins*, and represents the the amount - or relative *frequency*- of observations falling within each bin as the height of a bar. This simple yet powerful plot provides an immediate visual impression of the dataset's distribution, allowing one to identify symmetry, skewness, concentration of values, and potential gaps. For example, a symmetric histogram, like the one in LHS of Figure 1.1 suggests a roughly balanced distribution around the mean, while a right-skewed histogram, like the one displayed in the RHS of Figure 1.1, indicates that higher values are less frequent - or less *probable* - but can yet influence measures like the mean. This is the state of the art in physical sciences and whenever data is supposed to fit a mathematical prediction.

Building a histogram in an informative way is extremely powerful, and there are some subtleties to consider. As a rule of thumb, look for natural divisions in the data, and keep all bins the same size, covering the whole range under study. Outliers can skew, so they must be treated

carefully. Figures 1.1 and 1.1 show how the binning size can affect the distribution of data. Smaller binning leads to more resolution but can be easily distorted in the presence of outliers, while few large bins are robust agains though losing the accuracy in resolution. For skewed distributions it is normally better to use the median and the IQR.

The box plot, also known as the *box-and-whisker* plot, was introduced by John Tukey in 1970 as part of his work on exploratory data analysis [21]. The box plot offers a compact summary of a dataset's central tendency, spread, and potential outliers. Constructed from five key statistics - the minimum, first quartile (Q_1), median (Q_2), third quartile (Q_3), and maximum - it clearly shows the *interquartile range* (IQR $= Q_3 - Q_1$) and highlights points that fall outside 1.5 times the IQR as outliers. This representation allows for quick comparisons across multiple groups, and it is particularly useful for detecting asymmetry and variability without being overly influenced by extreme values. It is widely used in biological and clinical sciences where an experiment can be repeated many times with relatively small sizes. Figure 1.2 displays the same data represented as histograms in Figures 1.1 as box plots.

Finally, the *violin* plot is a more recent innovation, combining the box plot with a series of mathematical tools that represent as well the shape of the distribution. While the precise origin is less formally documented, it gained prominence in the late 20th century in statistical software environments, such as R, during the 1990s [13]. Essentially, the violin plot extends the concept of the box plot by combining it with a kernel density estimate of the data. This plot not only displays the median and quartiles but also provides a smooth depiction of the distribution's shape, revealing features such as multimodality or skewness that might be obscured in a simple box plot. By showing both summary statistics and the underlying density, the violin plot gives a richer, more nuanced view of the dataset, particularly when comparing several groups side by side. As an example of such comparison, see Figure [...].

Quote to Anscombe [1] and Spearman [20]

Exercises

1. Exercise [...].

2. Exercise [...].

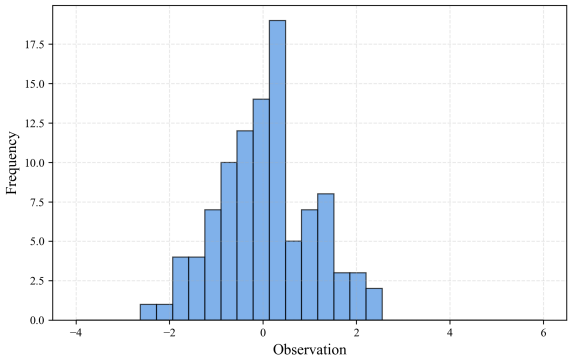
3. Exercise [...].

Solutions

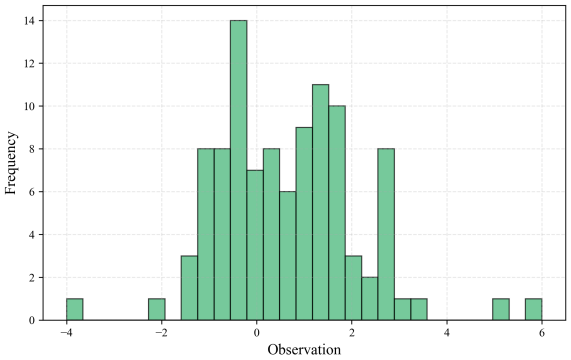
1. Solution [...].

2. Solution [...].

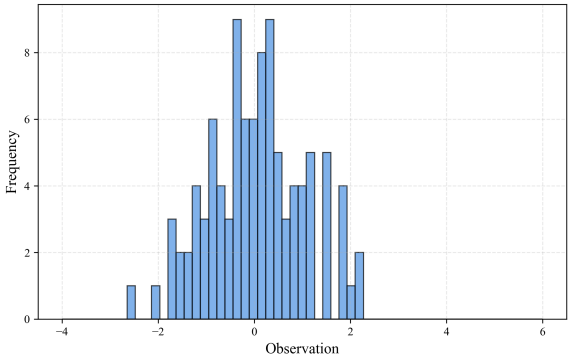
3. Solution [...].



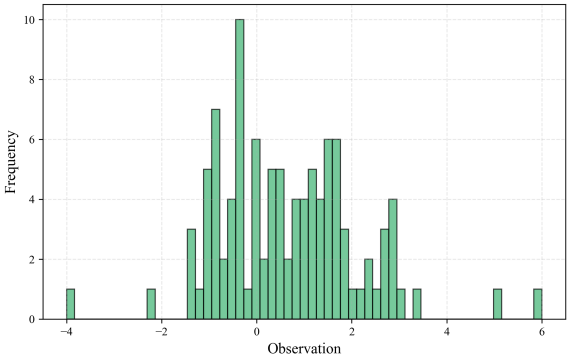
(a) Clean Gaussian distribution, no outliers.



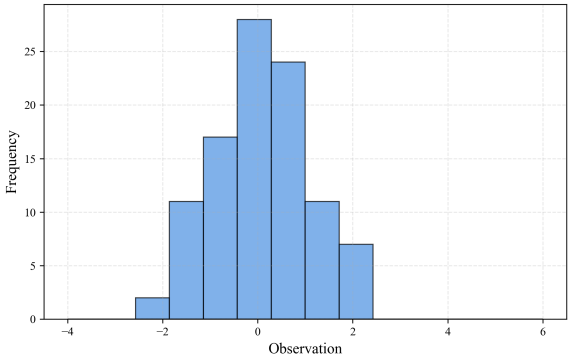
(b) Skewed Gaussian due to outlier data.



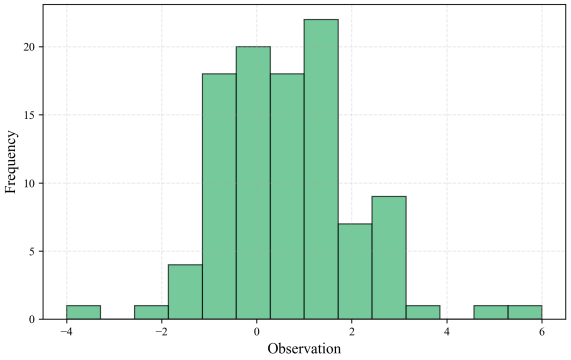
(c) Clean Gaussian (narrow bins).



(d) Skewed Gaussian (narrow bins).

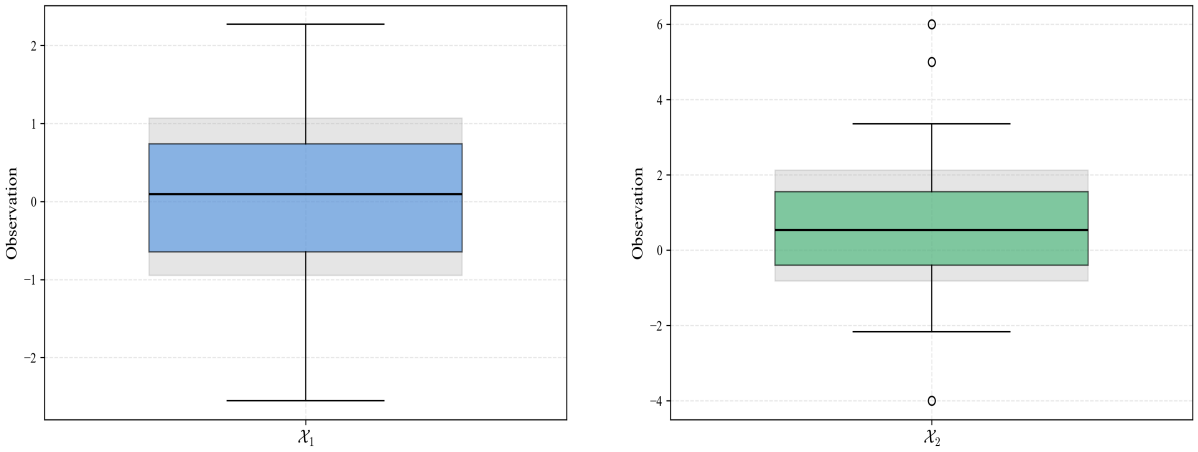


(e) Clean Gaussian (wide bins).



(f) Skewed Gaussian (wide bins).

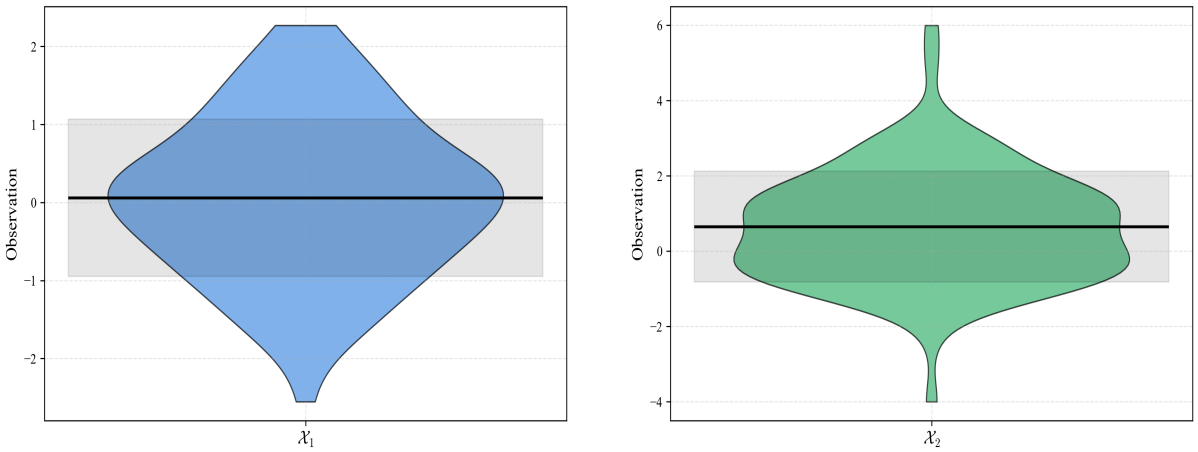
Figure 1.1: Comparison of histogram representations of $n = 100$ Gaussian observations under three binning conditions. Top row: standard binning; middle row: narrow bins (higher resolution, more sensitive to outliers); bottom row: wide bins (lower resolution, more robust to outliers). In each pair, the left panel shows a clean Gaussian sample, while the right panel shows a skewed, outlier-affected sample.



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

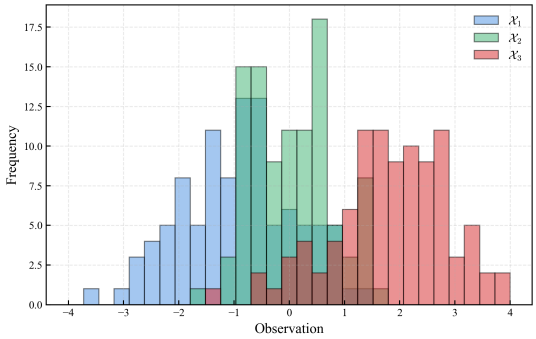
Figure 1.2: Box plots representing $n = 100$ observations drawn from a Gaussian distribution. The central black line shows the mean value, representing the central tendency where the bulk of events lie. The shadowed area highlights the standard deviation, as measure of the variability and spread the observations with respect to the mean



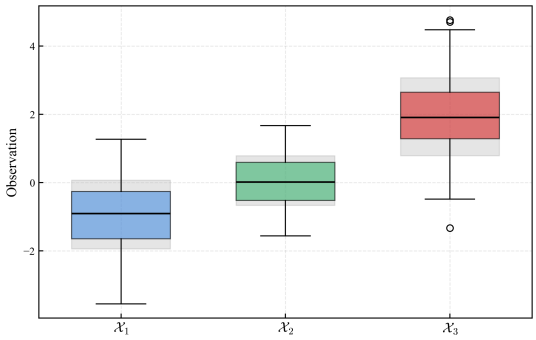
(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

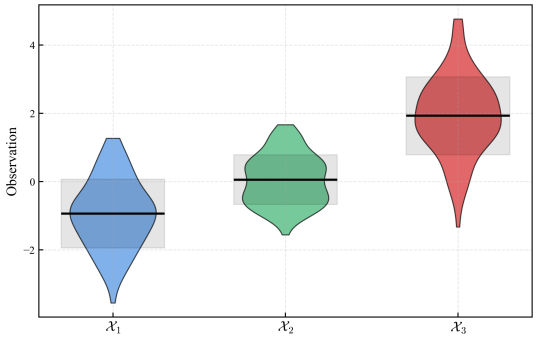
Figure 1.3: Violin plots representing $n = 100$ observations drawn from a Gaussian distribution. The central black line shows the mean value, representing the central tendency where the bulk of events lie. The shadowed area highlights the standard deviation, as measure of the variability and spread the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



(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 [15].

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

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

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

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

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

general case, where both outcomes can happen with equal probability, we would write

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

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

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

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

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

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

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

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

For a coin toss, this reduces to

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

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

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

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

These three notions suffice for now. We have seen probability as a number that quantifies uncertainty, the frequentist definition in terms of ratio, and the idea of unitarity. Let us emphasize, though, that this formulation is actually quite recent. Even though the basic intuitions were already introduce by Bernoulli and Laplace, as we discussed, it was not until the nineteenth and twentieth centuries, that probability theory was properly formalized in the language of analysis. The idea of expectation became formally defined via the Lebesgue integral [17], stochastic processes were studied by Wiener and Doob [10], and the idea of convergence - that

lied the foundations for unitarity - were explored by Borel, Cantelli, and Kolmogorov [4]. From gambling practice, probability grew into a highly abstract and powerful theory.

The way probability was mathematically defined is based on the idea of *probability space*. This may seem quite abstract at first, so let's illustrate with an example. Imagine that every experiment we perform - tossing a coin, rolling a dice, or measuring the brightness of a star - has a collection of possible outcomes. This collection is what mathematicians call the *sample space*, often written as Ω. Within this space, we may be interested in particular groups of outcomes, such as "getting an even number" from a dice roll or "obtaining heads" from a coin toss. These groups of outcomes are what we call *events*.

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

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

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

$$\mathbb{P}(A) \ \geq \ 0 \, , \qquad \forall A \in \mathcal{F} \, . \tag{2.3}$$

- **Normalisation:** The probability of the whole sample space Ω is exactly 1. This expresses the fact that "something will happen." If Ω is the complete list of possible outcomes of an experiment, then we are certain that the final outcome will be one of them.

$$\mathbb{P}(\Omega) \ = \ 1 \, . \tag{2.4}$$

- **Countable additivity:** If we have several events $A_1, A_2, \ldots$ that cannot overlap (e.g. they are mutually exclusive or disjoint), then the probability that one or another occurs is the sum of their probabilities. For instance, in rolling a die, the probability of rolling "1 or 2" is $P(1) + P(2)$ because the two outcomes cannot happen at the same time.

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \right) \ = \ \sum_{i=1}^{\infty} \mathbb{P}(A_i) \, , \qquad \text{for disjoint } A_i \in \mathcal{F} \tag{2.5}$$

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

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

Of course, mathematics provides structure, but actual meaning requires interpretation. As we mentioned already, two main schools of interpretation emerged from this axiomatic development of probability theory. The *frequentist* definition, associated with Richard von Mises [22], identifies probability as we did in (**??**), with long-run frequency in repeated trials: it is then an objective property of the physical world, revealed through repetition. On the other hand, the *Bayesian* tradition, with origins in Thomas Bayes' posthumous essay *Towards solving a Problem in the Doctrine of Chances* back in 1763 [2], conceives probability as a measure of rational belief, updated by evidence through what we call nowadays Bayes' rule, or Bayes' theorem. Such idea was later refined by Laplace [16] and further developed by Bruno de Finetti, who further emphasized that probability expresses degrees of personal belief coherent under rational rules of betting [8].

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

2.1 Discrete random variables

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

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

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

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Chapter 3

Estimation, prediction and inference

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 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.

Once we know the foundations of probability theory, we can make assumptions about the true popuplation parameters, through expected values [...].

Once we have some notions of descipive stastistics, we can make assumptions about the true popuplation parameters, through expected values [...].

Historical Note

Fisher (1922, 1925) connected least squares, likelihood, and sampling distributions, establishing the foundations of modern inference. Neyman (1937) formalized confidence intervals as frequentist procedures, contributing to philosophical debates on inference that continue today.

4.1 Prediction vs inference revisted

4.2 General approach to hypothesis testing

4.3 Statistical tests: some examples

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

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

4.3.3 Compare variation on two groups - Fisher's exact test

4.3.4 Compare variation o multiple groups - Fisher's ANOVA

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

4.4 Parametric and non-parametric tests

4.5 Comparing data and normalization

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Chapter 5

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

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

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

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

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

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

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

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

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

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

[13] J. L. Hintze and R. D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1997.

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

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

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

[17] Henri Lebesgue. *Intégrale, longueur, aire*. 1902.

[18] Karl Pearson. *The Grammar of Science*. Adam and Charles Black, London, 1892. Introduces the term and concept of histogram.

[19] 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.

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

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

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