Imperial College London

Introduction to probability theory and statistical inference

Jesús Urtasun Elizari

Research Computing and Data Science

November 20, 2025

Contents

Preface

This preface introduces the motivation, structure, and goals of the book.

Introduction

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

The purpose of these notes

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

Both, predictive probability and descriptive statistics have deep historical roots, from ancient works on chance and divination to modern scientific topics oriented towards information theory, modelling and data analysis. As one could guess, rivers of ink have been written about such topics, and endless literature sources are available. However, after following many different courses at both bachelor and postgraduate levels, and teaching such topics myself during the last three years, I have found that most resources belong, almost certainly, to one of the next three classes. Either (*i*) deeply mathematical, and hence out of reach for most experimental or clinically oriented scientists, (*ii*) laboratory oriented, focusing on inference and experimental design, and hence missing most of the mathematical background, or (*iii*) with a direct focus towards programming and computation, relying on domain specific notebooks (Python, R, Matlab, SPSS, etc), and online resources with precompiled libraries for simulation, which again miss most of the mathematical and formal intuitions. Indeed, the misuse of statistics in experimental sciences is a critical topic in modern times, as mathematicians have extensively discussed during the last decades. The well-known article by John P. A. Ioannidis, *"Why most published research findings are false"* [**ioannidis2005why**], serves as a prominent example, and it may serve as motivation for a rigorous study.

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

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

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

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

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

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

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

- A simple, intuitive introduction to statistics with few mathematical concepts is provided in Spiegelhalter's *"The Art of Statistics: How to Learn from Data"* [**spiegelhalter2019art**].

- A more foundational textbook, with more advanced mathematical approach, can be found at DeGroot and Schervish's *"Probability and Statistics"* [**degroot2012probability**].
- For a philosophical and historical perspective on probability and statistics, please find Forster and Bandyopadhyay's handbook *"Philosophy of Statistics"* [**bandyopadhyay2011philosophy**].
- A comprehensive introduction with focus on practical applications and modern data analysis tools is can be found at Diez, Barr & Mine *"OpenIntro Statistics"* [**openintro2025**].
- 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"* [**pishronik2014introduction**].

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 [**finkel2007ancient**]. 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"* [**david1962games**].

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 [**cicero45bce**]. 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"*) [**cardano1663ludo**], 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 [**devlin2008unfinished**]. These developments paved the way for later contributions by Christiaan Huygens, who in 1657 wrote the first published textbook on probability *"De Ratiociniis in Ludo Aleae"* (*"On Reasoning in Games of Chance"*), and Jacob Bernoulli, whose 1713 *"Ars Conjectandi"* (*"The Art of Conjecturing"*) remains among the most influential early texts in the field. Their works, alongside with many others, collectively laid the groundwork for the probabilistic and statistical methods that foreshadow modern scientific reasoning [**huygens1657ratiociniis**], [**bernoulli1713ars**], [**hald1990history**].

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"*) [**kolmogorov1933grundbegriffe**], Kolmogorov synthesized classical and frequentist ideas into a rigorous mathematical framework based on measure theory. His axioms remain the standard foundation for probability theory to this

day. It may seem surprising that a concept with such ancient origins was not formally axiomatized until relatively recent times, and we will return to Kolmogorov's formulation and its implications in greater detail in Chapter 5. Nevertheless, philosophical discussions about the interpretation of probability and its relation to the physical sciences - especially in the context of determinism, epistemology and modern topics such as quantum mechanics - predate Kolmogorov's formulation and continue to evolve to this day.

Chapter 1

Descriptive statistics

Statistics is the grammar of science.

— Karl Pearson

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

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

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

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

The mathematical formalization of decision-making is actually quite a recent development. It is usually attributed to British mathematician Frank P. Ramsey (1903–1930), who in his 1926 paper *Truth and Probability* [**ramsey1926**] 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.

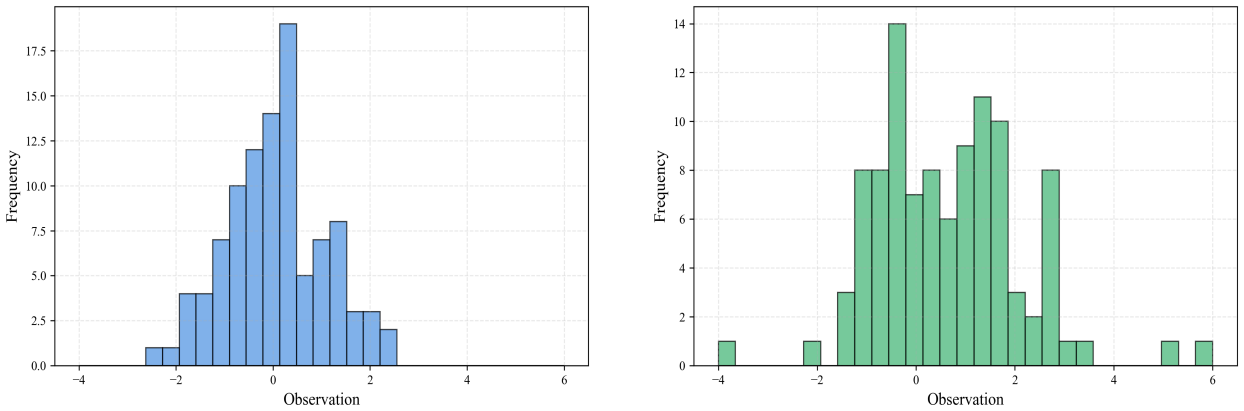
Equally important is the recognition that not all data are of the same kind. A common distinction is to consider *categorical* and *numerical* data. Categorical - or *qualitative* - data describes qualities or labels such as the eye colour of students in a classroom (blue, brown, green), the brand of 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.



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

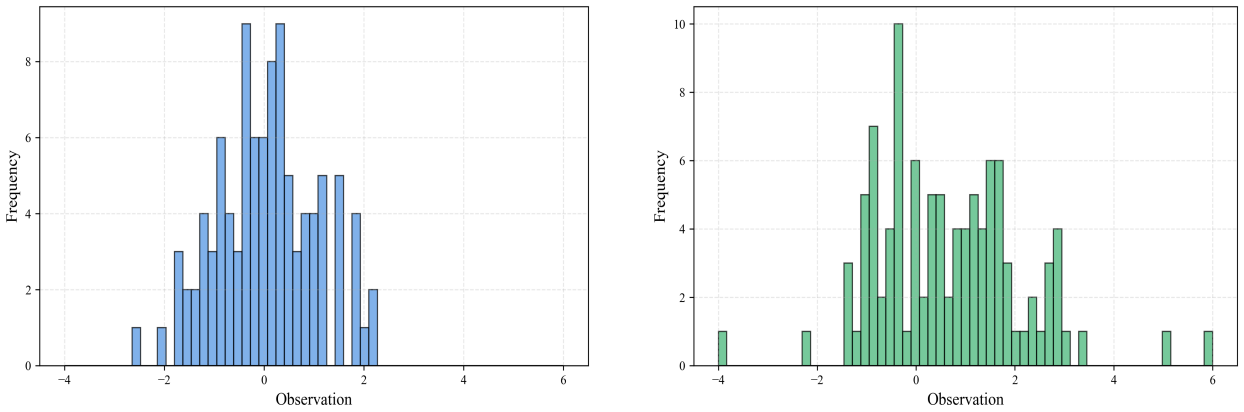
Figure 1.1: Histogram representation of $n = 100$ observations drawn from a Gaussian distribution. The RHS shows a clean Gaussian distribution, symmetric around the central value and with no outlier points, while the LHS shows a skewed version, where some outliers make the distribution deviate from the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

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:

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



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

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

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:

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

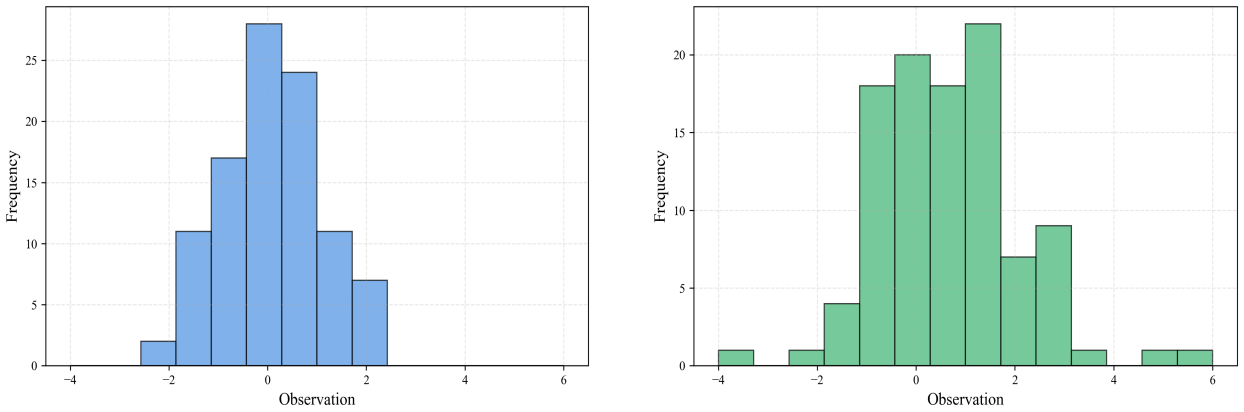
We can write this in a slightly more compact way as a *summation*, as follows:

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

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 (**??**) and (**??**) 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 (**??**), while for the full population of N elements $x_1, x_2, \ldots, x_N$ - before any sampling - the *population mean* is normally denoted as μ, and defined accordingly

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



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

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

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 (**??**) 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 (**??**) 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 (**??**), 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 **??** 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 (**??**) 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 (**??**) you should get the result

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

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

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

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

and for the entire population,

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

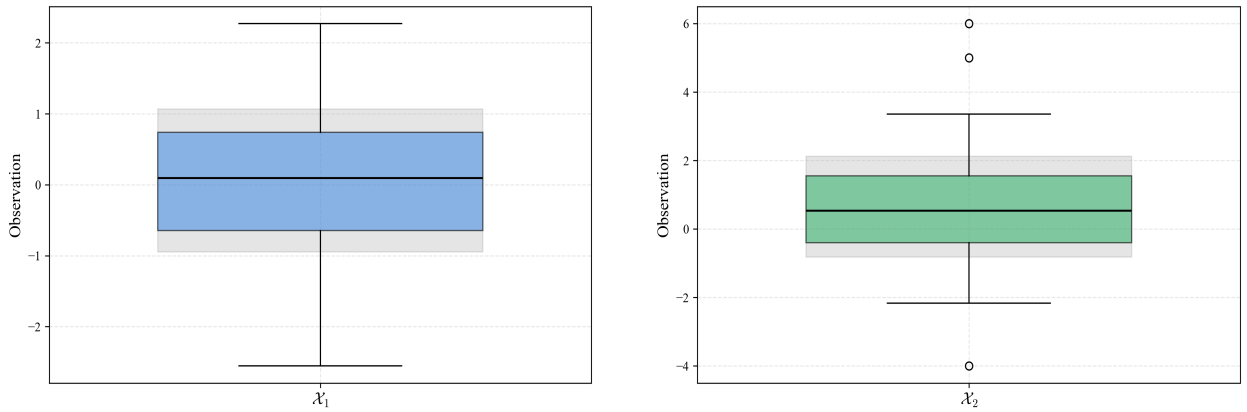
At a glance, variance and standard deviation quantify how much the elements of a dataset deviate from the mean, capturing the notion of *spread*.

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

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

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

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



(a) Clean Gaussian distribution, no outliers.

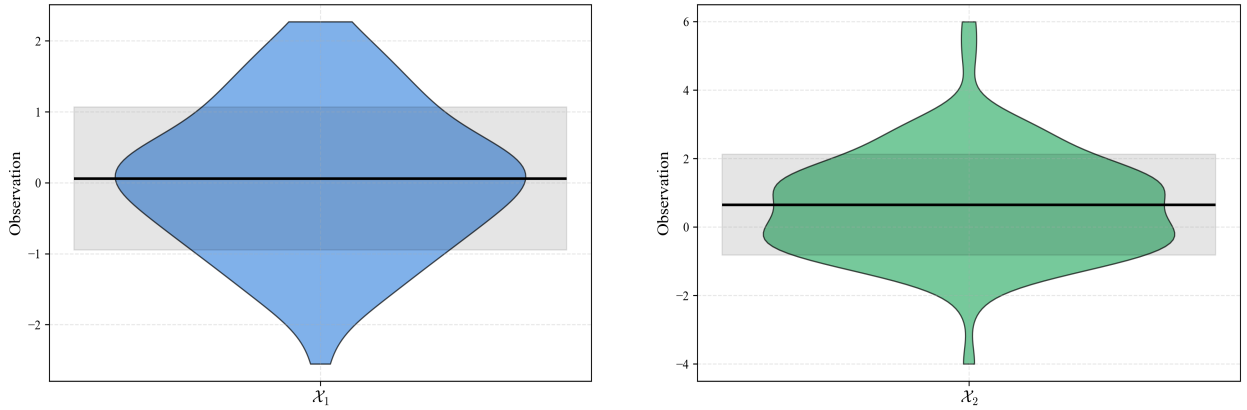
(b) Skewed Gaussian due to outlier data.

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

1.3 Data visualization

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

The *histogram* is found among the oldest and most fundamental visualization tools. The concept of dividing data into intervals to visualize frequency dates back to Karl Pearson in the late 19th century, who formalized it as a graphical representation of probability distributions [**pearson1892**]. 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 **??** suggests a roughly balanced



(a) Clean Gaussian distribution, no outliers.

(b) Skewed Gaussian due to outlier data.

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

distribution around the mean, while a right-skewed histogram, like the one displayed in the RHS of Figure **??**, 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 **??** and **??** 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 [**tukey1977**]. 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 **??** displays the same data represented as histograms in Figures **??** 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 [**hintze1997**]. Essentially, the violin plot extends the

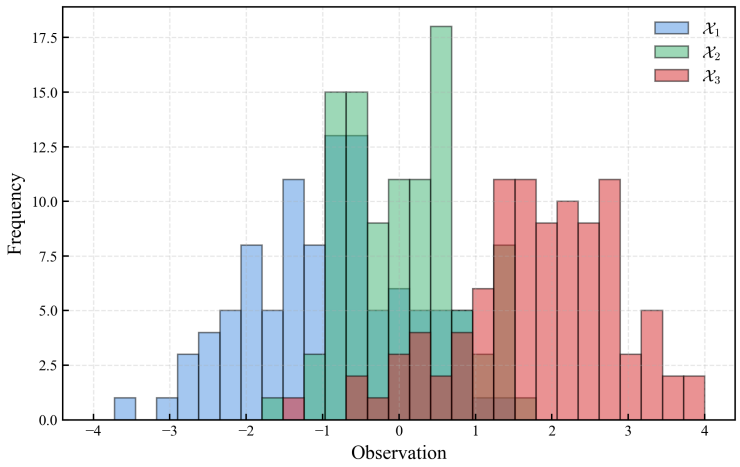


Figure 1.6: Three sets of observations, with the mean value and standard deviation represented as a violin plot [...]. The red line shows the mean value, representing the central tendency where the bulk of events lie. The shadowed area highlights the standard deviation, as measure of the variability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

concept of the box plot by combining it with a kernel density estimate of the data. This plot not only displays the median and quartiles but also provides a smooth depiction of the distribution's shape, revealing features such as multimodality or skewness that might be obscured in a simple box plot. By showing both summary statistics and the underlying density, the violin plot gives a richer, more nuanced view of the dataset, particularly when comparing several groups side by side. As an example of such comparison, see Figure [...].

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

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

1.4 Dependency, linearity, correlation

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

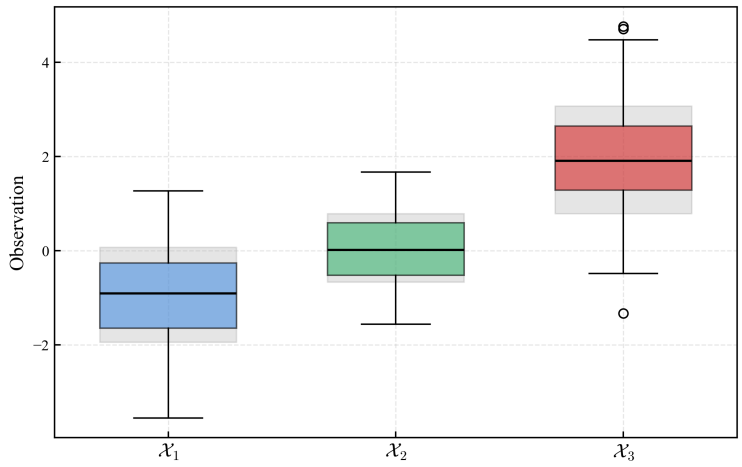


Figure 1.7: Three sets of observations, with the mean value and standard deviation represented as a box plot [...]. The red line shows the mean value, representing the central value where the bulk of events lie, and the shadowed area the standard deviation, as measure of the variability, or how spread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

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

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

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

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

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

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

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

Beyond polynomials, mathematical modelling admits much broader and richer forms, such as exponential growth and decay, logarithmic compression, trigonometric oscillations, and nonlinear

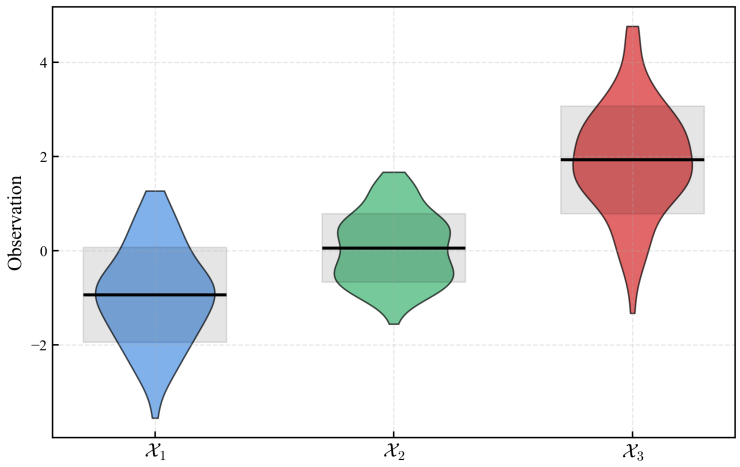


Figure 1.8: Three sets of observations, with the mean value and standard deviation represented as a violin plot [...]. The red line shows the mean value, representing the central value where the bulk of events lie, and the shadowed area the standard deviation, as measure of the variability, or how spread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

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

The idea of *correlation* is a fundamental measure of association between two random variables, quantifying how strongly they vary together. The most widely used mathematical description is the *Pearson correlation coefficient*, introduced by Karl Pearson in the 1890s, which is built upon the idea of covariance normalized by variability. For two random variables x and y, the population correlation $\rho_{x,y}$ is defined as

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

where the covariance is defined as

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

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

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

The computation of expected values is not such a trivial topic. In next chapter we will revisit this concept with more detail, in the context of random variables and probability distributions. For

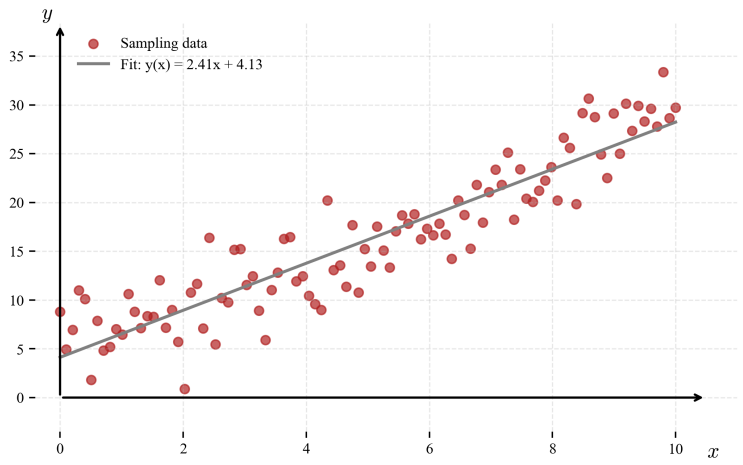


Figure 1.9: Scatter data following linear dependency. The linear function $y(x) = ax + b$ is displayed overlaying the sampling points, with parameters fitted from data.

now, and for practical purposes with finite datasets, we can approximate the expected value just by the *arithmetic mean*, or *sample mean* of the observations, $\bar{x}$ and $\bar{y}$.

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

Let's illustrate with an example. Consider the dataset

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

The sample means are trivial to compute

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

and the covariance is then

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

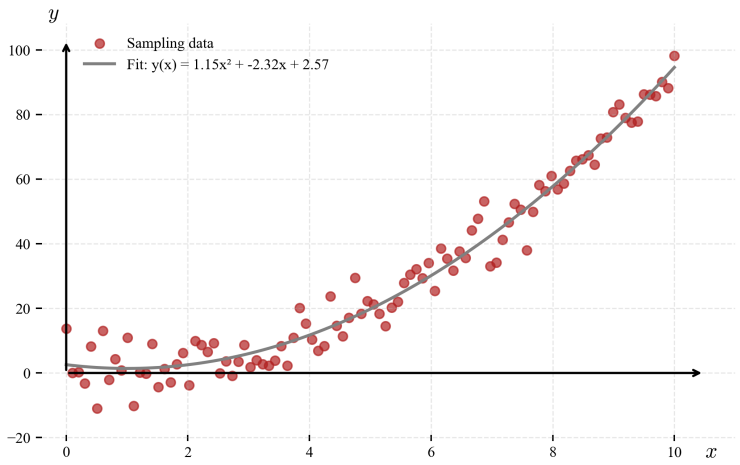


Figure 1.10: Scatter data following quadratic dependency. The quadratic function $y(x) = ax + b$ is displayed overlaying the sampling points, with parameters fitted from data.

The standard deviations are

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

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

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

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

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

A note about notation: the Pearson correlation coefficient is usually denoted r for a finite sample and ρ for the idealized, complete population. Interpretation of r is quite straightforward for linear

relationships: values close to ±1 indicate a strong linear association, while values near 0 suggest weak or no linear dependency. However, Pearson's correlation has limitations: it is sensitive to outliers, captures only linear relationships, and can be misleading if the relationship is more complex.

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

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

where ε denotes the error term, sometimes called the *residual*. A classic demonstration of Pearson's limitations is *Anscombe's quartet*, created by the statistician Francis Anscombe in 1973 [**anscombe1973graphs**]. It consists of four datasets with nearly identical summary statistics—including means, variances, and correlation coefficients—yet exhibiting dramatically different distributions and patterns when plotted. This striking example highlights the necessity of visualizing data rather than relying solely on numeric summaries.

There are ways of defining correlation in more subtle cases, but they lie outside the scope of this course. Just as an example for ordinal data, where rankings rather than numeric differences matter, the *Spearman rank correlation* ρ_s is more appropriate definition. It is often used to assess monotonic relationships, was introduced by Charles Spearman in 1904 [**spearman1904**]. It is based on the ranks of the data rather than the raw values, making it less sensitive to outliers and non-linear relationships.

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

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

Exercises

1. Exercise [...].

2. Exercise [...].

3. Exercise [...].

Solutions

1. Solution [...].

2. Solution [...].

3. Solution [...].

Chapter 2

Foundations of Probability

2.1 Sample Spaces

A sample space Ω represents all possible outcomes.

Chapter 3

Prediction and inference

3.1 Sample Spaces

A sample space Ω represents all possible outcomes.

Chapter 4

Introduction to hypothesis testing

4.1 Sample Spaces

A sample space Ω represents all possible outcomes.

Chapter 5

Introduction to conditional probaility

5.1 Sample Spaces

A sample space Ω represents all possible outcomes.

Chapter A

Additional Results

Here you place additional proofs, tables, or extended material.