Introduction to probability theory and statistical inference

Jesús Urtasun Elizari

Research Computing and Data Science

April 4, 2025

Contents

Index **v**

1 Introduction **1**

2 Introduction to probability and random events **5**
- 2.1 What is probability? . . . 5
- 2.2 Discrete and continuous . . . 6
- 2.3 Probability distributions . . . 8
  - 2.3.1 Binomial distribution . . . 8
  - 2.3.2 Poisson distribution . . . 10
  - 2.3.3 Uniform distribution . . . 12
  - 2.3.4 Gaussian distribution . . . 14
  - 2.3.5 Exponential distribution . . . 15

3 Parameter estimation **17**
- 3.1 Prediction vs inference . . . 17
- 3.2 Parameter estimation . . . 18
- 3.3 Law of Large Numbers (LLN) . . . 19
  - 3.3.1 Definition . . . 19
  - 3.3.2 Intuition . . . 19
  - 3.3.3 Types of LLN . . . 19
- 3.4 Central Limit Theorem (CLT) . . . 21
  - 3.4.1 Definition . . . 21
  - 3.4.2 Intuition . . . 21
  - 3.4.3 Applications . . . 21

4 Introduction to statistical inference **23**
- 4.1 Prediction vs inference . . . 23
- 4.2 Hypothesis testing . . . 23
- 4.3 Statistic tests, p-values and significance . . . 24
  - 4.3.1 Compare sample mean with hypothesized value - One sample t-test . . . 24
  - 4.3.2 Compare sample means of two groups - Two sample t-test . . . 26
  - 4.3.3 Compare sample variances of two groups - Fisher test . . . 28
  - 4.3.4 Compare more than two groups - ANOVA . . . 30
  - 4.3.5 Compare distributions - χ^2 test . . . 32
- 4.4 Parametric and non-parametric . . . 34
- 4.5 Comparing data and normalization . . . 35

5 Introduction to bayesian statistics **37**
- 5.1 The Bayes' theorem . . . 37
- 5.2 Bayesian vs frequentist . . . 37
- 5.3 Bayesian statistics . . . 37

6 Stochasticity and Markov processes 39
6.1 Stochasticity and Markov processes 39
6.2 Markov chains and hidden Markov models 39
6.3 Modern applications 39

Index

Chapter 1

Introduction

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 inference and hypothesis testing. The text is composed by six chapters, together with some appendix reviewing basic mathematical concepts, 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 or mathematical training.

First, we will introduce the concept of probability itself, and we will discuss basic ideas on how to model information and chance. Then we will discuss a series of mathematical approaches to such quantities and formally define random processes, also referred to as *stochastic*. We will introduce the idea of a function and how functions need to be adapted to implement uncertainty when discussing random events. In the second part we will address the difference between prediction and inference, and discuss a set of subjects normally grouped under the name of *hypothesis testing*. Here we will introduce how to quantify certainty and bias, how to model significance and the idea of hypothesis tests. Finally, we will briefly discuss more modern topics, such as bayesian statistics, stochasticity and Markov processes.

Probability theory is the branch of mathematics that aims to quantify uncertainty, chance and information in the so-called and random events. It provides the foundation for understanding and modelling various real-world phenomena, ranging from gambling and statistical inference to machine learning and quantum mechanics. In the same way we learn to count and measure, we could try to assign numerical values to the likelihood of different outcomes in an experiment, or quantify the level of certainty - or *surprise* for such unknown result. The modern approach to probability and its fundamental concepts are summarized in the axioms established by the russian mathematician Andrey Kolmogorov, in the early 1930s. Some people may find surprising that such an old topic was not properly formalized until such recent times. We will cover these concepts in Chapter 1, but first, let's look a bit further in history to explore where the intuitions about chance and information came to appear.

The idea of stochasticity and randomness has deep historical roots. Ancient civilizations, including the Babylonians, Egyptians, and Greeks, grappled with the concept of uncertainty in games of chance, commerce, and divination. The oldest known dice date back over 5,000 years, indicating an early human fascination with randomness. While these cultures did not develop formal mathematical probability, they recognized patterns in random events and attempted to predict outcomes based on empirical observations and superstitions.

Greek philosophers such as Democritus and Aristotle debated the nature of chance and determinism[...]. The Roman philosopher Cicero distinguished between chance events and those governed by fate, foreshadowing later discussions on probability. It would not be until medieval times, where scholars like Gerolamo Cardano (1501–1576) made early contributions by analyzing gambling problems and laying the groundwork for probability theory.

The first formalization of probability as a mathematical discipline began in the 17th century with the correspondence between Blaise Pascal and Pierre de Fermat, who devised combinatorial methods to solve problems related to games of chance. Their work introduced fundamental ideas such as expected value and laid the foundation for later advances by Christiaan Huygens, Jacob Bernoulli, and Abraham de Moivre. Bernoulli's *Ars Conjectandi* (1713) introduced the Law of Large Numbers, establishing that observed frequencies converge to theoretical probabilities over many trials, a topic we will extensively cover in Chapter 2.

As we will cover in the first chapter, the main object of study in probability and chance are the so-called *stochastic* process. Indeed, the word stochastic comes from no other than the greek word στοχαστικός, which literally means *to guess*. By stochastic, we just mean a process, or a system, that evolves randomly over time. Unlike deterministic processes, where the outcome is fixed by initial conditions, stochastic processes incorporate uncertainty at each step. The reader could anticipate here that almost *any* process can be modelled as a stochastic process, from the brownian motion of particles in physics, to signal processing in engineering, population behaviours, and stock price fluctuations in finance. A stochastic process will be defined as a collection of random variables indexed by time or space, commonly represented as x. The probability of a variable x to take a specific value x_0 when measured will be given by $P(x = x_0)$

One of the most fundamental stochastic processes is the Markov process, which exhibits the Markov property: the future state depends only on the present state, not on past history. This property makes Markov chains particularly useful in modelling systems with memoryless transitions, such as queueing networks and genetic sequences. Another significant class of stochastic processes is the Poisson process, which describes the occurrence of rare events over time, such as radioactive decay or network traffic.

Information theory, developed by Claude Shannon, is deeply intertwined with probability and stochastic processes. It quantifies the amount of uncertainty or surprise in a system and provides a mathematical foundation for data compression and communication. The fundamental measure in information theory is entropy, defined as:

$$H(X) = -\sum_i P(x_i) \log P(x_i), \tag{1.1}$$

where $P(x_i)$ is the probability of observing outcome x_i. Entropy captures the unpredictability of a source of information, with higher entropy indicating greater uncertainty.

Shannon also introduced the concept of mutual information, which measures the reduction in uncertainty about one random variable given knowledge of another. This concept is crucial in signal processing, machine learning, and cryptography, where efficient information transfer and encoding are essential.

Together, probability theory, stochastic processes, and information theory form a powerful toolkit for modelling and analyzing uncertainty. Their applications span numerous disciplines, from physics and biology to artificial intelligence and cybersecurity. Understanding these concepts allows for the development of efficient algorithms, accurate predictions, and optimized decision-making strategies in uncertain environments.

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 is not the symbols or the numbers, but the intuitions and the general understanding, what we are behind. Let's start with an example of quantities most students are already familiar with, and for which they may have some intuition, but that are not always properly introduced at the mathematical level. Let's illustrate with an example how to properly define the *mean* and *variance* of a set of observations.

Imagine we are doing an experiment and we measure some variable x (position, energy, concentration, ...). We repeat the measurement three times and we get first 1, then 2, and the last time 3. We will write our $x = 1, 2, 3$. We define the *mean* - or *average* - $\bar{x}$as the sum of all element and divided by the total.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.2}$$

Where we denote the sum of all elements with the greek letter $\sum$, starting with the first one $i = 1$ and until the last one $i = N$. If we now substitute that expression for our set of $N = 3$ observations

$$\bar{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{1}{3}(1 + 2 + 3) = 2 \tag{1.3}$$

In a similar way, we can define the *variance* as a quantity that captures how far are the elements of the set from the mean value. Here are three important textbooks in the field of probability and statistics:

$$\bar{x} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{1.4}$$

Where we denote the sum of all elements with the greek letter $\sum$, starting with the first one $i = 1$ and until the last one $i = N$. If we now substitute that expression for our set of $N = 3$ observations

$$\bar{x} = \frac{1}{3-1}\sum_{i=1}^{3}(x_i - \bar{x})^2 = \frac{1}{2}\big((1-2)^2 + (2-2)^2 + (2-3)^2\big) = \frac{1}{2}(1 + 0 + 1) = 1 \tag{1.5}$$

- A simple, intuitive introduction to statistics with few mathematical concepts is provided in Spiegelhalter's *The Art of Statistics: How to Learn from Data* [1].
- A more foundational textbook, with more advanced mathematical approach, can be found at DeGroot and Schervish's *Probability and Statistics* [2].
- For a philosophical and historical perspective on probability and statistics, pleas find McFadden's *The Philosophy of Statistics* [3].

Chapter 2

Introduction to probability and random events

2.1 What is probability?

As already mentioned in the introduction, probability theory is one of the oldest subjects within mathematical studies. Ideas related to chance and random events, together with measurement, information, inference, etc, can be traced back to very ancient times. It could sound surprising, though, that almost every topic explained nowadays in modern courses of statistics is extremely new, ranging broadly a century. Concepts like distributions, gaussian behaviour, p-values, hypothesis testing or normalization, were introduced and formalized in the XVIII, XIX and XXth century by mathematicians such as Pearson and Fisher (...).

The first question we could ask ourselves, is what actually *are* random events? What do we mean by *probability*? Since the times of Cardano and Gauss [...], there have been many attempts to explore and describe these, but we will keep it simple. For a more technical introduction to these topics, please see [...]. Broadly speaking, probability, and later statistical inference, are branches of mathematics dealing with chance, also referred to as *random* events, or *stochastic* processes. By random, or stochastic, we will mean simply a process whose output we *ignore*, and as examples we could think of tossing coins, rolling dice, or performing some arbitrary measurement. Indeed, the word stochastic comes from no other than the greek word στοχαστικός, which literally means *to guess.* Let's try to briefly introduce the idea of probability, as a quantity that allows as to describe such events.

So let's now ask ourselves the following question. What *is* probability in the first place? What do we mean by it and what does it describe? Probability is nothing more, and nothing less, that a number we make up, a quantity we come up with, to quantify certainty in a process whose outcome we ignore. A number we will use to describe the amount of information we have about a random - or stochastic - event. For simplicity, we can make it range from 0 to 1, in the following way.

- If I'm sure A will never happen, $P(A) = 0$.
- If I'm sure A will always happen, $P(A) = 1$.
- For anything in between, $P(A) \in [0, 1]$.

With the symbol $\in$ we simply denote that $P(A)$ will be a number between 0 and 1. It could also be read "$P(A)$ is *contained* in the interval $[0, 1]$". In all those cases where we are not sure it we will get one result or another, we say that there is a level of *uncertainty*, or *surprise.* Let's think on a coins toss, as an example. To model such case, the simplest example of a stochastic process, we would have two possible outcomes. Heads (H), and tails (T)

- If I'm sure I will get heads, $P(H) = 0$, and $P(T) = 0$.

- If I'm sure I will get heads, $P(H) = 1$, and $P(0) = 0$.
- For anything in between, $P(H) = P(T) = \frac{1}{2}$.

Let's stop here for a moment. When we say that the probability of getting heads - or tails - in a normal coin - a fair one, that is not biased, we are implicitly assuming some things. We implicitly assume that If we repeated the toss many times, half of them we would get one result, and the other half the remaining result. This is normally referred to as the *frequentist* definition of probability, because we are defining its value as the ratio of how many times we get a specific result, and the number of total possible outcomes.

$$P(\text{A happening}) = \frac{\text{Number of times A happens}}{\text{Total number of trials}} \tag{2.1}$$

Which in the case of the coin, if I toss 100 times, and obtain 55 heads against 45 tails, would lead to

$$P(\text{H}) = \frac{55}{100} =\simeq \frac{1}{2} \tag{2.2}$$

Ideally we expect that these frequencies, as we increase the number of repetitions, would approach a perfect $frac12$. We will revisit this concept when we talk about the Law of Large Number and the Central Limit Theorem, in Chapter 2

But this is not the only thing we assume about such a quantity. For probabilities to represent the real behaviour of random processes and information, they must follow another property, called *unitarity.* Unitarity ensures that, if we consider and add up the probabilities for all possible events in a given experiment, we get the total. That means, at least one of the scenarios will happen.

We will denote all possible outcomes of an experiment x_1, x_2, ..., x_n. And we will require that the sum of probabilities of all possible outcomes add up to 1. This is a crucial property we will refer as *normalization*

$$\sum_{i=0}^{n} p(x_i) = 1 \tag{2.3}$$

Indeed, the literal meaning of probability comes from latin probabilis. American logician and philosopher Richard Jeffrey, "Before the middle of the seventeenth century, the term "probable" (Latin probabilis) meant just approvable, and was applied in that sense, univocally, to opinion and to action. A probable action or opinion was one such as sensible people would undertake or hold, in the circumstances."[12] However, in legal contexts especially, "probable" could also apply to propositions for which there was good evidence.

The sixteenth-century Italian polymath Girolamo Cardano demonstrated the efficacy of defining odds as the ratio of favourable to unfavourable outcomes (which implies that the probability of an event is given by the ratio of favourable outcomes to the total number of possible outcomes[14]). Aside from the elementary work by Cardano, the doctrine of probabilities dates to the correspondence of Pierre de Fermat and Blaise Pascal (1654). Christiaan Huygens (1657) gave the earliest known scientific treatment of the subject.[15] Jakob Bernoulli's Ars Conjectandi (posthumous, 1713) and Abraham de Moivre's Doctrine of Chances (1718) treated the subject as a branch of mathematics.[16] See Ian Hacking's The Emergence of Probability[10] and James Franklin's The Science of Conjecture[17] for histories of the early development of the very concept of mathematical probability.

Like other theories, the theory of probability is a representation of its concepts in formal terms – that is, in terms that can be considered separately from their meaning. These formal terms are manipulated by the rules of mathematics and logic, and any results are interpreted or translated back into the problem domain.

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

2.2 Discrete and continuous

Once we have an insight on random events, and a mathematical quantity representing that uncertainty, we are ready to deal with real problems. From tossing coins, to rolling dice, to making measurements, the first thing we realize is that not *all* random events are equal. In some cases, like rolling a fair dice, all outcomes are equally probable, and in other cases, such as counting, we may encounter somre results which happen much more often than others. The main criteria we will use for differentiate among random events, is what we will call their *distribution*.

We will distinguish two main families of random events. These in which the number of possible outcomes is finite, or *countable*, and the ones where the number of outcomes is *uncountable*. The first ones will be named as *discrete* events, while the second are normally referred to as *continuous*.

2.3 Probability distributions

So far we have introduced the idea of random events, and the concept of probability as a number to quantify surprise. For our present chapter, we will try to model such stochastic events such that we can make predictions. For that purpose, we will model that probability we just defined to be a descriptive, or *predictive* quantity. Let's begin by saying that not all random phenomena are equal. Hence, a basic way to classify and separate random events, is according to how their probabilities are *distributed*.

2.3.1 Binomial distribution

The simplest case of random event we will describe are the so-called *binomial* events. Cases where we make a certain number of measurements n, each with two or more possible outcomes, and we want to know the number of successes. For instance, what would be the probability of measuring, or observing, 5 heads if I toss 10 coins? Or what would be the probability of obtaining 5 times a 6, out of a total of 100 dice rolls? In all these cases we will call x the number of successes we want to observe, n the total number of trials, and p the probability of success in each individual trial. The binomial distribution models the number of successes in a fixed number of independent trials, each with the same probability of success. It was developed by Jacob Bernoulli in the 17th century while studying the probability of repeated Bernoulli trials. His work laid the foundation for the Law of Large Numbers.

Intuitively, this distribution is useful when considering repeated experiments with two possible outcomes (success or failure). For example, flipping a fair coin multiple times follows a binomial pattern. We will say that the probability of observing x successes in n total tries, given individual probability of success p, is given by:

$$P(x; n, k) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{2.4}$$

This is normally referred to as a probability *mass* distribution. The reason for that, as we will discuss later, is to distinguish such events from other types of events called continuous, for which we will define density distributions. For now, just keep probability mass distribution as a fancy name, or probability distribution, for simplicity. Let's break this expression down in a couple of examples.

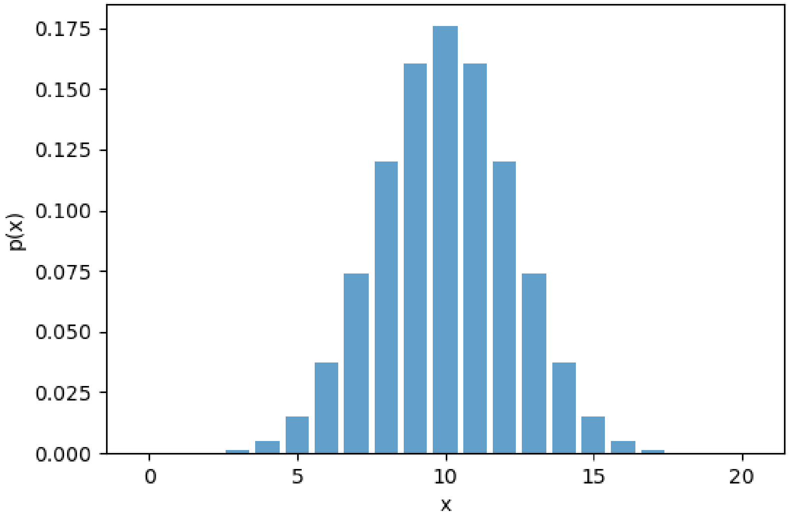

Figure 2.1: Representation of the binomial distribution of a random variable x, given the total number of trials n and the individual probability of success p.

Example 1: Suppose we flip a fair coin 5 times ($n = 5$) and want to find the probability of getting exactly 3 heads ($p = 0.5$):

$$\begin{aligned} P\left(x = 5; n = 10; p = \frac{1}{2}\right) = &\binom{5}{3}\left(\frac{1}{2}\right)^3\left(1 - \frac{1}{2}\right)^2 \\ &\binom{5}{3}\left(\frac{1}{2}\right)^3\left(1 - \frac{1}{2}\right)^2 \times 0.25 = 0.3125. \end{aligned}$$

Example 2: Suppose we flip a fair coin 5 times ($n = 5$) and want to find the probability of getting exactly 3 heads ($p = 0.5$):

$$P\left(x = 5; n = 10; p = \frac{1}{3}\right) = \binom{5}{3}\left(\frac{1}{2}\right)^3\left(1 - \frac{1}{2}\right)^2 = 10 \times 0.125 \times 0.25 = 0.3125.$$

Example 3: Suppose we flip a fair coin 5 times ($n = 5$) and want to find the probability of getting exactly 3 heads ($p = 0.5$):

$$P\left(x = 5; n = 10; p = \frac{1}{6}\right) = \binom{5}{3}\left(\frac{1}{2}\right)^3\left(1 - \frac{1}{2}\right)^2 = 10 \times 0.125 \times 0.25 = 0.3125.$$

2.3.2 Poisson distribution

The next kind of random event we will discuss are the *Poisson* distributed, named after the french mathematician Siméon Denis Poisson, who tried to model to events that were random but with a known average rate, such as the number of people crossing a street per day, or the number of customers entering a store, or emails received per hour. As a note, this distribution was introduced in quite recent times, in the early 19th century to model rare events. It is particularly useful for counting occurrences over a fixed interval of time or space.

The probability mass function for observing a number of events x if we know the average rate λ is:

$$P(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \tag{2.5}$$

Again, let's consider a couple of examples to illustrate Poisson distributed events.

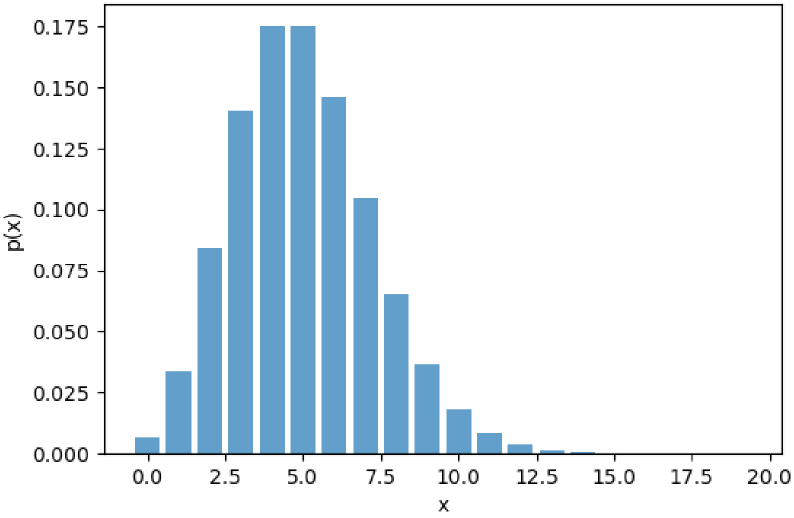

Figure 2.2: Representation of the Poisson distribution of a random variable x, given the number of observations λ as a parameter.

Example 1: We would like to know the probability of observing exactly 5 cancer patients in a hospital over a week, if we know the average number ($\lambda = 3$) patients per week.

$$P(x=5;\lambda=3) = \frac{3^5 e^{-3}}{5!} = \frac{243e^{-3}}{120} \approx 0.1008.$$

Example 2: Let's now ask a similar, but different question. So far, we have only focused on the probability of observing *exactly* one particular outcome. But we could ask as well, what would be the probability observing 5 *or less* cancer patients in that same hospital ($\lambda = 3$) patients per week.

$$\begin{aligned} P(x\leq 5;\lambda=3) = {} & P(x=0;\lambda=3) + P(x=1;\lambda=3) + P(x=2;\lambda=3) \\ & + P(x=3;\lambda=3) + P(x=4;\lambda=3) + P(x=5;\lambda=3) \end{aligned}$$

This sum of probabilities up to a given value is normally referred to as the *cumulative probability*, *cumulative distribution function*, or cdf.

Example 3: Now let's ask the opposite question. What would be the probability of observing *at least* 5 patients in that same hospital?

$$P(x = 5; \lambda = 3) = \frac{3^5 e^{-3}}{5!} = \frac{243e^{-3}}{120} \approx 0.1008.$$

2.3.3 Uniform distribution

The last example of such discrete random - remember that by discrete we mean *countable* number of outcomes - will be those which are *uniformly* distributed. That means that all possible outcomes are equally probable, such as rolling a fair die, or tossing a coin.

The uniform distribution represents a scenario where all outcomes in an interval $[a, b]$ are equally likely. This distribution has been used since antiquity, especially in early probability and gambling studies.

The probability of observing a particular result x in a given range $[a, b]$ is:

$$f(x; a, b) = \frac{1}{b-a}, \quad a \leq x \leq b. \tag{2.6}$$

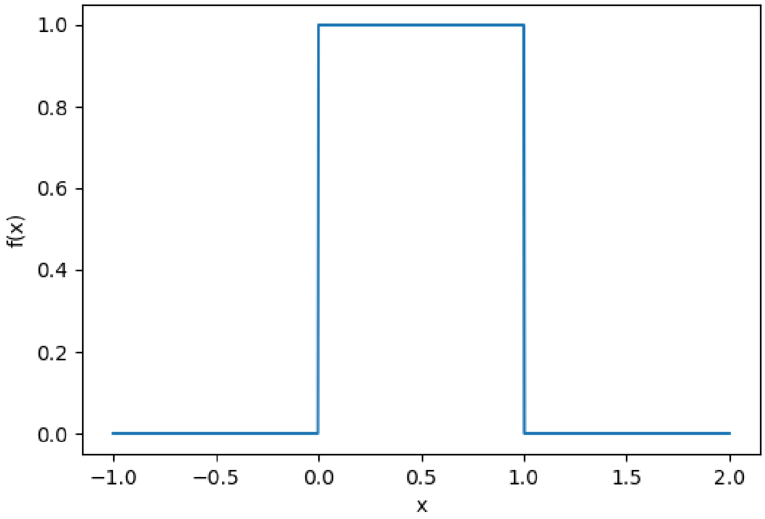

Figure 2.3: Representation of the uniform distribution of a random variable x, given the boundary parameters a, b.

Example 1: If a random number is chosen from the interval [2, 10], the probability density is:

$$f(x; a, b) = \frac{1}{10-2} = 0.125. \tag{2.7}$$

Example 2: If a random number is chosen from the interval [2, 10], the probability density is:

$$f(x; a, b) = \frac{1}{10-2} = 0.125. \tag{2.8}$$

Example 3: If a random number is chosen from the interval [2, 10], the probability density is:

$$f(x; a, b) = \frac{1}{10-2} = 0.125. \tag{2.9}$$

So far we have focused on discrete events, that is, scenarios where the number of possible outcomes was an integer number. Now we will encounter a second family of stochastic processes, the ones we will refer to as continuous. In the discrete case, we were implicitly using the frequentist definition of probability, as a number that represents the ratio of how many times we will observe a particular result, if we endlessly repeat (...).

But let's try to face a different scenario. What would happen if we try to guess the probability of measuring something which does hace an infinite number of possible outcomes, spread on a continuous range? (e.g., the probability of measuring the height of a person an get 1.75 cm, or the temperature in a room and get 25 degrees, ...). Here we notice that, if we keep the definition of probability we used in the case of the Binomial, the Poisson, etc, we would get something like:

$$P(x = x_0) = \frac{\text{number of times I get } x_0}{\text{number of times I get any other result}} \tag{2.10}$$

Note that now, the possible results are not just 1, 2, ..., n, but actually infinite more and spread over a continuous range. The outcome of measuring a temperature could be the $T = 25$ we want, but also $T = 24.999$ and $T = 25.001$, and there infinite other possible results between these two. No matter how precise our measurement devices, are, between any pair of results, we would have an infinite number of cases where we obtain a different result. Hence, applying the frequentist definition of probability would lead to:

$$P(x = x_0) = \frac{\text{number of times I get } x_0}{\text{number of times I get any other result}} = \frac{n}{\infty} = 0 \tag{2.11}$$

We would get that the probability of obtaining *any result* would be exactly zero.

Let's pause for a moment and think about what happened. At the very beginning of this chapter we said that the quantity $P(x)$ was used to represent information - also certainty, surprise - and computed using the frequentist approach, meaning the *ratio of favorable cases and total cases.* But that was assuming we had a finite set or possibilities, or measure space.

- Discrete (coins, dice, counting) $\longrightarrow$ finite, *countable* outcomes
- Continuous (temperature, energy, concentration, ...) $\longrightarrow$ infinite, *uncountable*

For such cases we will define a mathematical quantity, similar to that we called probability, which represents analogous information, but considering the fact we are dealing with a continuous event. We will call it *probability density* or *density* for simplicity, and we will denote it with $f(x)$. Note that we can distinguish it from the probability in discrete events $P(x_i)$, where we used the subscript x_i to represent that the random variable could take just a finite set of values (x_1, $x2$, etc).

- Discrete (coins, dice, counting) $\longrightarrow$ Probability $P(x_i)$ — $\sum_{i=1}^{\infty} P(x_i) = 1$
- Continuous (temperature, energy, concentration, ...) $\longrightarrow$ Probability density f(x) — $\int_{i=0}^{\infty} f(x)dx = 1$

In the same way we imposed that probability needs to obey unitarity, we will impose that property in our recently defined probability density $f(x)$. The way we represent the sum for all possible cases in the continuous case, is just imposing that the integral of the function $f(x)$ is 1. This is just an example of *normalization*, that we will explore further in chapter 4.

2.3.4 Gaussian distribution

Introduced by Carl Friedrich Gauss, the normal distribution became central to statistics due to the Central Limit Theorem (CLT). It describes how averages of large samples tend to form a bell-shaped curve.

Intuitively, many natural and social phenomena follow a normal distribution, such as human heights and test scores.

Introduced by Carl Friedrich Gauss, the normal distribution became central to statistics due to the Central Limit Theorem (CLT). It describes how averages of large samples tend to form a bell-shaped curve.

Intuitively, many natural and social phenomena follow a normal distribution, such as human heights and test scores.

The probability density function is:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.12}$$

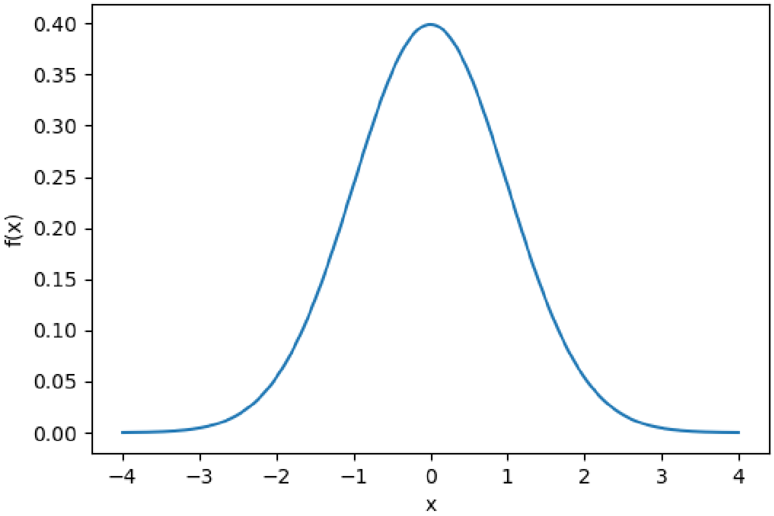

Figure 2.4: Representation of the gaussian distribution of a random variable x, given the mean value μ and standard deviation σ parameters.

Example 1: If human heights are normally distributed with mean $\mu = 170$ cm and standard deviation $\sigma = 10$ cm, the probability density of someone being exactly 180 cm is:

$$f(180) = \frac{1}{10\sqrt{2\pi}} e^{-\frac{(180-170)^2}{2(10)^2}} \approx 0.0242. \tag{2.13}$$

Example 2: If human heights are normally distributed with mean $\mu = 170$ cm and standard deviation $\sigma = 10$ cm, the probability density of someone being exactly 180 cm is:

$$f(180) = \frac{1}{10\sqrt{2\pi}} e^{-\frac{(180-170)^2}{2(10)^2}} \approx 0.0242. \tag{2.14}$$

Example 3: If human heights are normally distributed with mean $\mu = 170$ cm and standard deviation $\sigma = 10$ cm, the probability density of someone being exactly 180 cm is:

$$f(180) = \frac{1}{10\sqrt{2\pi}} e^{-\frac{(180-170)^2}{2(10)^2}} \approx 0.0242. \tag{2.15}$$

2.3.5 Exponential distribution

The exponential distribution models waiting times between Poisson process events. It has been widely applied in reliability analysis and survival studies.

Intuitively, it describes situations where the probability of waiting a certain time between events remains constant, such as time between bus arrivals.

The exponential distribution models waiting times between Poisson process events. It has been widely applied in reliability analysis and survival studies.

Intuitively, it describes situations where the probability of waiting a certain time between events remains constant, such as time between bus arrivals.

The probability density function is:

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

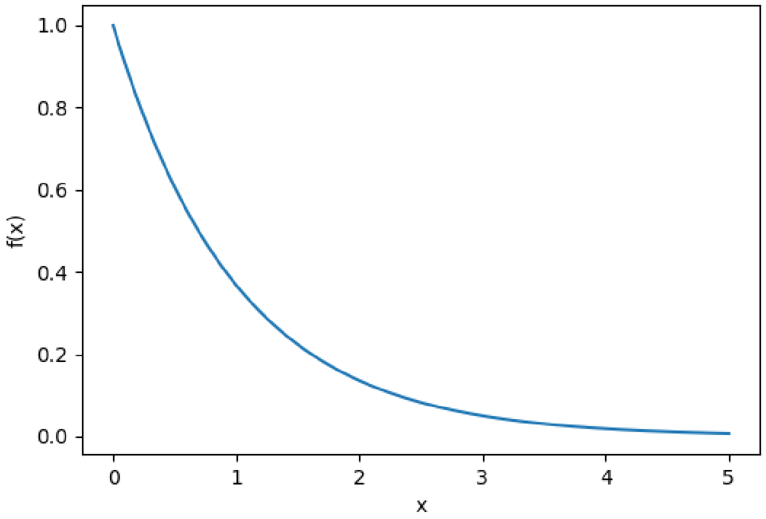

Figure 2.5: Representation of the exponential distribution of a random variable x, given the decay rate λ.

Example 1: If a call center receives calls at an average rate of $\lambda = 2$ per minute, the probability that the next call arrives after more than 2 minutes is:

$$P(X > 2) = e^{-2(2)} = e^{-4} \approx 0.0183. \tag{2.17}$$

Example 2: If a call center receives calls at an average rate of $\lambda = 2$ per minute, the probability that the next call arrives after more than 2 minutes is:

$$P(X > 2) = e^{-2(2)} = e^{-4} \approx 0.0183. \tag{2.18}$$

Example 3: If a call center receives calls at an average rate of $\lambda = 2$ per minute, the probability that the next call arrives after more than 2 minutes is:

$$P(X > 2) = e^{-2(2)} = e^{-4} \approx 0.0183. \tag{2.19}$$

Chapter 3

Parameter estimation

3.1 Prediction vs inference

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

The difference between prediction and inference has been a topic of interest in statistics and data science for centuries. While both concepts involve drawing conclusions from data, their goals, methodologies, and historical development differ significantly.

The roots of inference trace back to classical statistics, particularly the work of Pierre-Simon Laplace (1749–1827) and Carl Friedrich Gauss (1777–1855), who developed probability theory and the method of least squares. Their work laid the foundation for statistical inference, which aims to understand relationships between variables and make generalizable conclusions about populations from samples.

For example, Laplace used probability theory to estimate the population of France, introducing Bayesian inference, which provides a framework for updating beliefs based on observed data. Gauss contributed the normal distribution and least squares estimation, which became essential for making inferences about unknown parameters.

On the other hand, prediction as a primary goal gained traction much later, particularly in the 20th century with the rise of machine learning. The focus shifted from understanding relationships to optimizing models that generalize well to unseen data. In 2001, Leo Breiman, in his seminal paper "Statistical Modeling: The Two Cultures," highlighted the distinction, arguing that traditional statistics emphasized inference, whereas modern machine learning prioritized prediction.

Inference seeks to explain why and how variables relate. The key idea is causality and interpretability: given a dataset, inference aims to answer questions such as: Does smoking cause lung cancer, or is the correlation due to other confounding factors? How does an increase in temperature affect ice cream sales? What are the most significant predictors of house prices? Statistical techniques such as hypothesis testing, confidence intervals, and regression analysis help researchers understand these relationships. The emphasis is on estimating parameters and determining statistical significance rather than simply making accurate predictions. A classic example is Sir Ronald Fisher (1890–1962), who developed maximum likelihood estimation (MLE) to infer parameters of probability distributions. Fisher's work allowed statisticians to estimate relationships between variables and quantify uncertainty.

Prediction focuses on accuracy and generalization rather than explaining causality. The goal is to create a model that performs well on new, unseen data, even if the underlying relationships between variables are not fully understood. For example, in modern deep learning, neural networks can recognize faces with high accuracy but offer little interpretability in how they make decisions. Unlike inference, which aims to understand why a pattern exists, prediction is about making the best possible guess given the available data. A key contributor to predictive modeling was Andrey Kolmogorov (1903–1987), who formalized probability theory and stochastic processes. His work laid the foundation for modern probabilistic models used in machine learning and artificial intelligence. Later, Vapnik and Chervonenkis (1971) developed statistical learning theory, which introduced VC dimension, a measure of a model's ability to generalize. This work led to the development of Support Vector Machines (SVMs), a key predictive tool in machine learning.

While Breiman's paper emphasized the divide between the two approaches, modern data science integrates both. In many cases, understanding causal relationships (inference) can improve predictive performance, and accurate prediction can guide further investigations into causality. For example, epidemiologists use inference to understand how diseases spread, but they also use predictive models to forecast outbreaks. In finance, analysts infer factors affecting stock prices while using machine learning to predict market trends. By combining both approaches, we gain a more comprehensive understanding of data, allowing for both insightful interpretations and powerful predictions.

3.2 Parameter estimation

Another key difference we will discuss now, and quite a subtle one from the mathematical perspective, is that one between a *variable* and a *parameter*. Consider the example of a binomial experiment, e.g. tossing coins and asking for the probability of measuring a specific number of heads. There, we would write it as

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

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

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

3.3 Law of Large Numbers (LLN)

Introduction

The Law of Large Numbers (LLN) is one of the fundamental theorems of probability theory. It was first formulated by Jacob Bernoulli in the late 17th century and later refined by other mathematicians, including Pafnuty Chebyshev. Bernoulli's work aimed to formalize how relative frequencies of events stabilize as the number of trials increases, providing the foundation for statistical inference. LLN plays a crucial role in statistics, finance, and machine learning, ensuring that averages computed from large samples are reliable estimates of expected values.

3.3.1 Definition

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

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

3.3.2 Intuition

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

3.3.3 Types of LLN

- **Weak Law of Large Numbers (WLLN)**: Convergence in probability, i.e., for any $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| \geq \epsilon) \to 0 \quad \text{as } n \to \infty. \tag{3.3}$$

- **Strong Law of Large Numbers (SLLN)**: Almost sure convergence, i.e.,

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1. \tag{3.4}$$

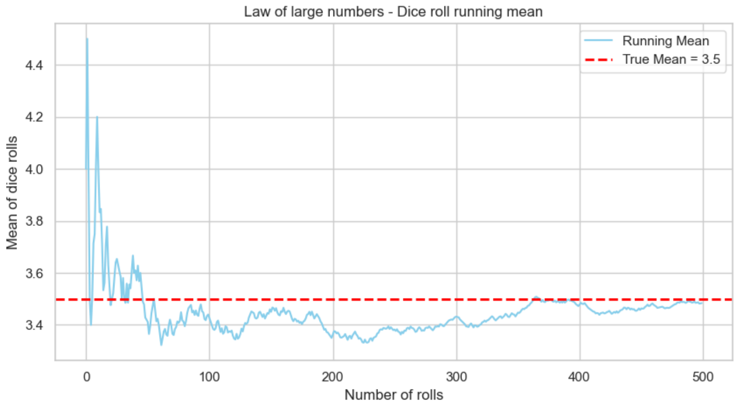

Figure 3.1: Representation of the law or large numbers. The sample mean tends to the population mean as the number of rolls n increases.

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

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

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

3.4 Central Limit Theorem (CLT)

Introduction

The Central Limit Theorem (CLT) was first discovered in the 18th century by Abraham de Moivre and later developed by Pierre-Simon Laplace and Carl Friedrich Gauss. It formalizes the idea that the distribution of sample means tends toward a normal distribution, regardless of the shape of the original population distribution. The CLT is fundamental in inferential statistics, allowing researchers to make predictions and construct confidence intervals for population parameters based on sample data.

3.4.1 Definition

The Central Limit Theorem states that for a large enough sample size, the sampling distribution of the sample mean follows a normal distribution, regardless of the original population distribution. Formally, if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ and variance σ^2, then the standardized sample mean:

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \tag{3.6}$$

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

3.4.2 Intuition

No matter the shape of the original distribution, when we take many samples and compute their means, the histogram of these sample means will resemble a normal curve as the sample size grows.

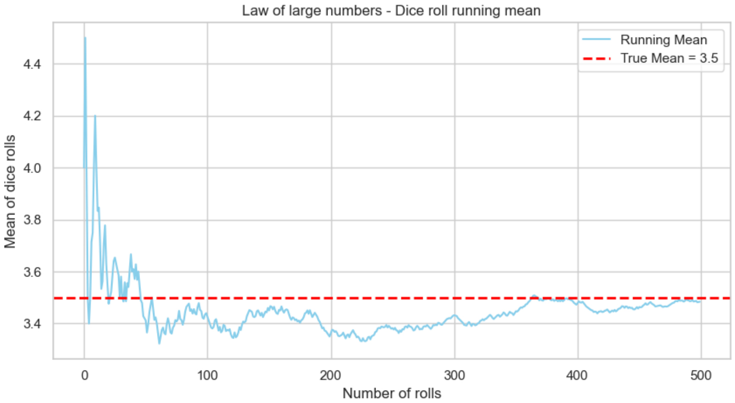

Figure 3.2: Representation of the law or large numbers. The sample mean follows a gaussian distribution as the sample size n increases.

Example: Consider rolling a fair six-sided die multiple times and computing the average outcome for groups of n rolls. As n increases, the distribution of these sample means approaches a normal distribution, centered at $\mu = 3.5$.

3.4.3 Applications

- Used in inferential statistics to approximate sampling distributions.
- Forms the basis for hypothesis testing and confidence intervals.
- Justifies the normality assumption in many statistical models.

Chapter 4

Introduction to statistical inference

4.1 Prediction vs inference

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

4.2 Hypothesis testing

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

So let's first ask ourselves the question. What is probability in the first place? What do we mean by it and what does it describe? Probability is nothing more, and nothing less, that a number we make up, a quantity we come up with, to quantify certainty. A number we will use to describe the amount of information we have about a random, or stochastic, event. For simplicity, we can make it range from 0 to 1, in the following way.

4.3 Statistic tests, p-values and significance

Statistic tests, p-values and significance

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

Let's begin with the simplest example of hypothesis testing we can think of. The so-called *one-sample t-test* is used to determine whether the mean of a single sample differs significantly from a known or hypothesized population mean. It assumes that the sample data are drawn from a normally distributed population. The test statistic is given by:

$$t = \frac{\bar{x} - \mu}{\sqrt{n}},$$

where $\bar{x}$ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size.

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

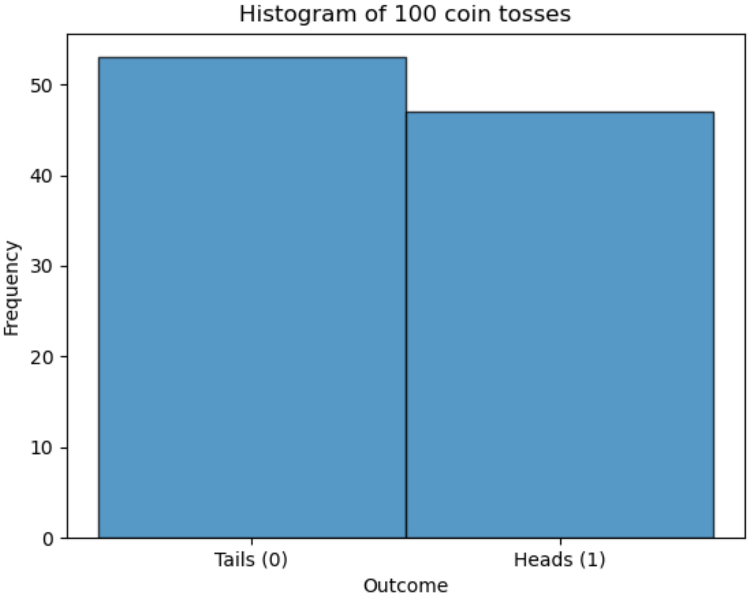

Figure 4.1: Observations following gaussian distribution.

Inference seeks to explain why and how variables relate. The key idea is causality and interpretability: given a dataset, inference aims to answer questions such as: Does smoking cause lung cancer, or is the correlation due to other confounding factors? How does an increase in temperature affect ice cream sales? What are the most significant predictors of house prices? Statistical techniques such as hypothesis testing, confidence intervals, and regression analysis help researchers understand these relationships. The emphasis is on estimating parameters and determining statistical significance rather than simply making accurate predictions. A classic example is Sir Ronald Fisher (1890–1962), who developed maximum likelihood estimation (MLE) to infer parameters of probability distributions. Fisher's work allowed statisticians to estimate relationships between variables and quantify uncertainty.

Prediction focuses on accuracy and generalization rather than explaining causality. The goal is to create a model that performs well on new, unseen data, even if the underlying relationships between variables are not fully understood. For example, in modern deep learning, neural networks can recognize faces with high accuracy but offer little interpretability in how they make decisions. Unlike inference, which aims to understand why a pattern exists, prediction is about making the best possible guess given the available data. A key contributor to predictive modeling was Andrey Kolmogorov (1903–1987), who formalized probability theory and stochastic processes. His work laid the foundation for modern probabilistic models used in machine learning and artificial intelligence. Later, Vapnik and Chervonenkis (1971) developed statistical learning theory, which introduced VC dimension, a measure of a model's ability to generalize. This work led to the development of Support Vector Machines (SVMs), a key predictive tool in machine learning.

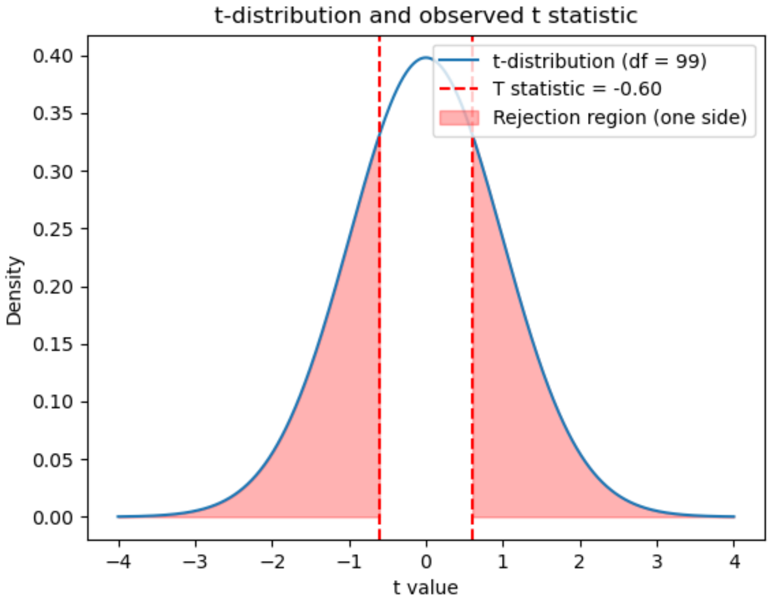

Figure 4.2: Representation of the Student's t distribution for given t_{obs} value.

While Breiman's paper emphasized the divide between the two approaches, modern data science integrates both. In many cases, understanding causal relationships (inference) can improve predictive performance, and accurate prediction can guide further investigations into causality. For example, epidemiologists use inference to understand how diseases spread, but they also use predictive models to forecast outbreaks. In finance, analysts infer factors affecting stock prices while using machine learning to predict market trends. By combining both approaches, we gain a more comprehensive understanding of data, allowing for both insightful interpretations and powerful predictions.

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

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

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{(n_1 - 1)^2 + (n_2 - 1)^2)}},$$

where $\bar{x}$ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size.

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

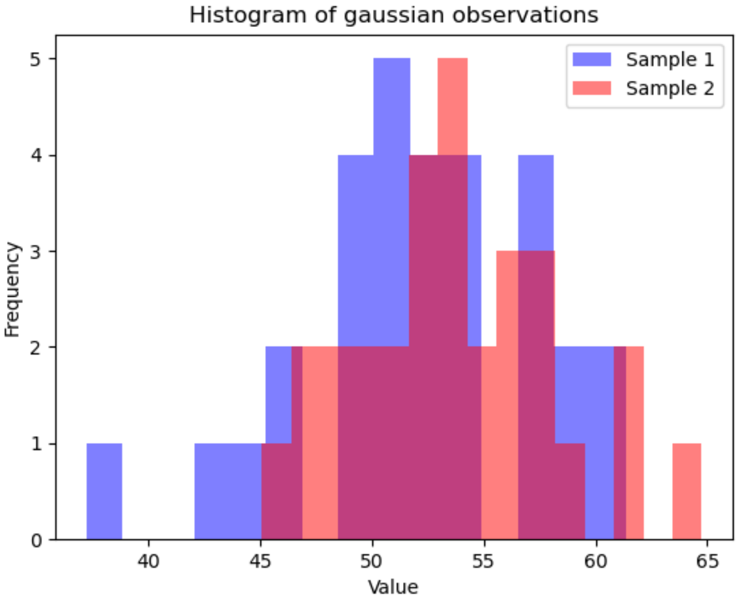

Figure 4.3: Observations following gaussian distribution.

Inference seeks to explain why and how variables relate. The key idea is causality and interpretability: given a dataset, inference aims to answer questions such as: Does smoking cause lung cancer, or is the correlation due to other confounding factors? How does an increase in temperature affect ice cream sales? What are the most significant predictors of house prices? Statistical techniques such as hypothesis testing, confidence intervals, and regression analysis help researchers understand these relationships. The emphasis is on estimating parameters and determining statistical significance rather than simply making accurate predictions. A classic example is Sir Ronald Fisher (1890–1962), who developed maximum likelihood estimation (MLE) to infer parameters of probability distributions. Fisher's work allowed statisticians to estimate relationships between variables and quantify uncertainty.

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

are not fully understood. For example, in modern deep learning, neural networks can recognize faces with high accuracy but offer little interpretability in how they make decisions. Unlike inference, which aims to understand why a pattern exists, prediction is about making the best possible guess given the available data. A key contributor to predictive modeling was Andrey Kolmogorov (1903–1987), who formalized probability theory and stochastic processes. His work laid the foundation for modern probabilistic models used in machine learning and artificial intelligence. Later, Vapnik and Chervonenkis (1971) developed statistical learning theory, which introduced VC dimension, a measure of a model's ability to generalize. This work led to the development of Support Vector Machines (SVMs), a key predictive tool in machine learning.

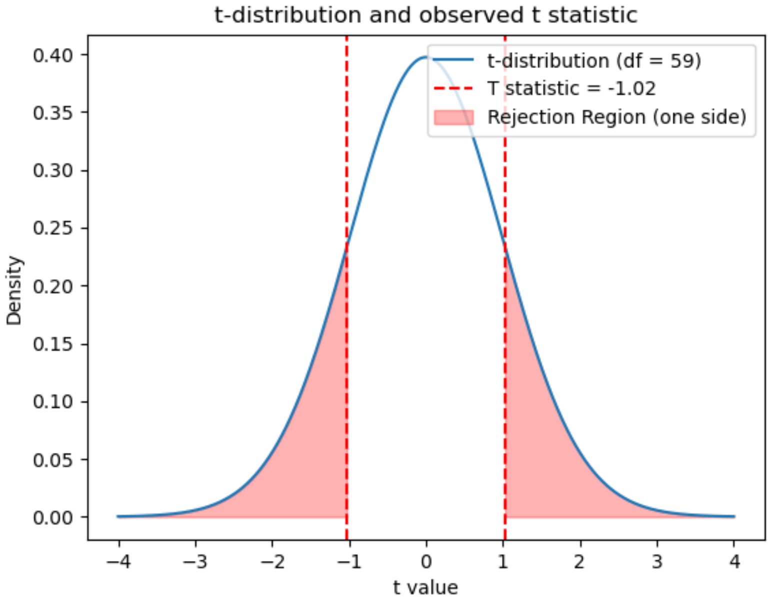

Figure 4.4: Representation of the Student's t distribution for given t_{obs} value.

While Breiman's paper emphasized the divide between the two approaches, modern data science integrates both. In many cases, understanding causal relationships (inference) can improve predictive performance, and accurate prediction can guide further investigations into causality. For example, epidemiologists use inference to understand how diseases spread, but they also use predictive models to forecast outbreaks. In finance, analysts infer factors affecting stock prices while using machine learning to predict market trends. By combining both approaches, we gain a more comprehensive understanding of data, allowing for both insightful interpretations and powerful predictions.

4.3.3 Compare sample variances of two groups - Fisher test

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

$$f(x; d_1, d_2) = \frac{s_1}{s_2},$$

where $\bar{x}$ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size.

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

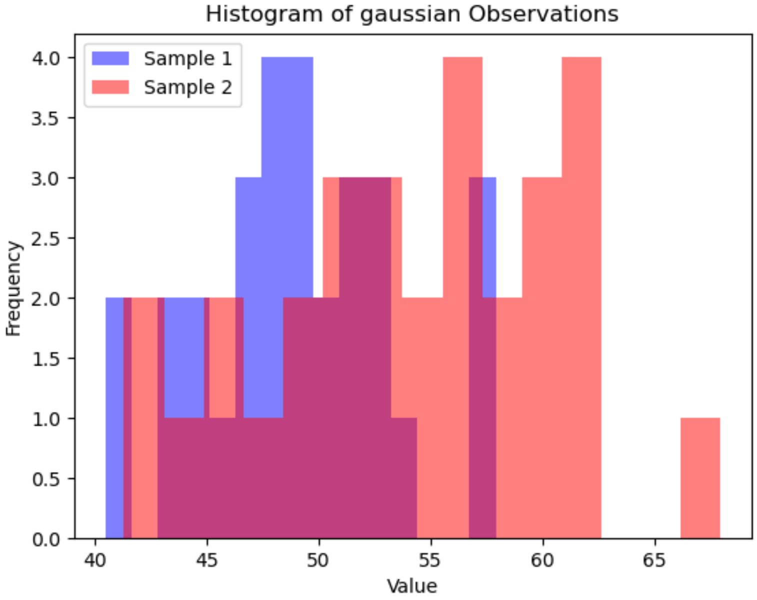

Figure 4.5: Observations following gaussian distribution.

Inference seeks to explain why and how variables relate. The key idea is causality and interpretability: given a dataset, inference aims to answer questions such as: Does smoking cause lung cancer, or is the correlation due to other confounding factors? How does an increase in temperature affect ice cream sales? What are the most significant predictors of house prices? Statistical techniques such as hypothesis testing, confidence intervals, and regression analysis help researchers understand these relationships. The emphasis is on estimating parameters and determining statistical significance rather than simply making accurate predictions. A classic example is Sir Ronald Fisher (1890–1962), who developed maximum likelihood estimation (MLE) to infer parameters of probability distributions. Fisher's work allowed statisticians to estimate relationships between variables and quantify uncertainty.

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

are not fully understood. For example, in modern deep learning, neural networks can recognize faces with high accuracy but offer little interpretability in how they make decisions. Unlike inference, which aims to understand why a pattern exists, prediction is about making the best possible guess given the available data. A key contributor to predictive modeling was Andrey Kolmogorov (1903–1987), who formalized probability theory and stochastic processes. His work laid the foundation for modern probabilistic models used in machine learning and artificial intelligence. Later, Vapnik and Chervonenkis (1971) developed statistical learning theory, which introduced VC dimension, a measure of a model's ability to generalize. This work led to the development of Support Vector Machines (SVMs), a key predictive tool in machine learning.

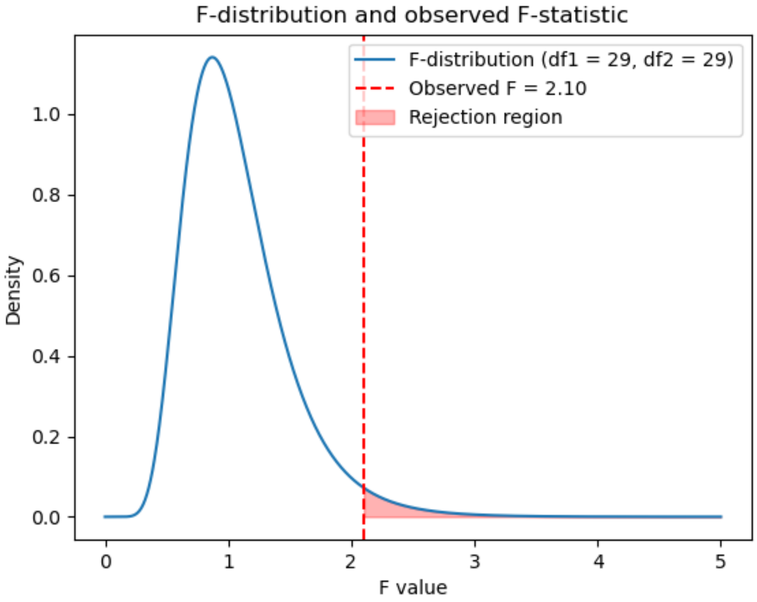

Figure 4.6: Representation of the Fisher distribution for given F_{obs} value.

While Breiman's paper emphasized the divide between the two approaches, modern data science integrates both. In many cases, understanding causal relationships (inference) can improve predictive performance, and accurate prediction can guide further investigations into causality. For example, epidemiologists use inference to understand how diseases spread, but they also use predictive models to forecast outbreaks. In finance, analysts infer factors affecting stock prices while using machine learning to predict market trends. By combining both approaches, we gain a more comprehensive understanding of data, allowing for both insightful interpretations and powerful predictions.

4.3.4 Compare more than two groups - ANOVA

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

$$f(x; d_1, d_2) = \frac{s_1}{s_2},$$

where $\bar{x}$ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size.

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

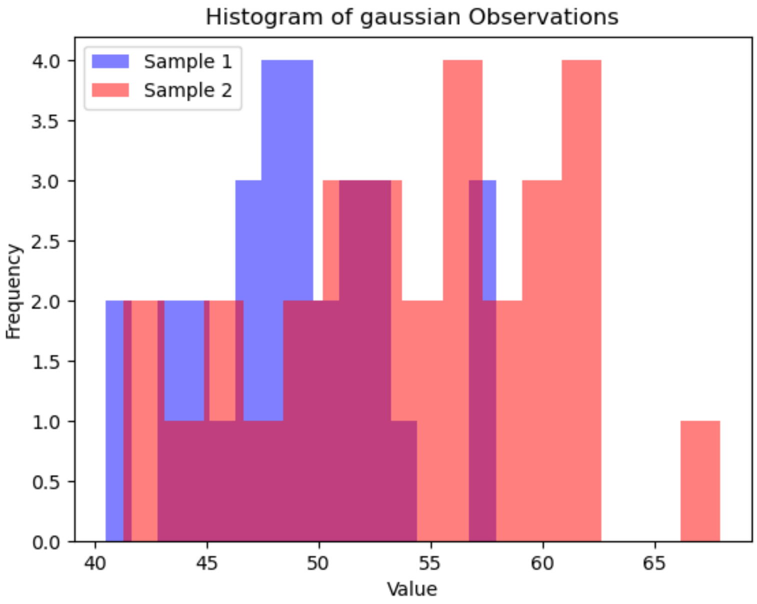

Figure 4.7: Observations following gaussian distribution.

Inference seeks to explain why and how variables relate. The key idea is causality and interpretability: given a dataset, inference aims to answer questions such as: Does smoking cause lung cancer, or is the correlation due to other confounding factors? How does an increase in temperature affect ice cream sales? What are the most significant predictors of house prices? Statistical techniques such as hypothesis testing, confidence intervals, and regression analysis help researchers understand these relationships. The emphasis is on estimating parameters and determining statistical significance rather than simply making accurate predictions. A classic example is Sir Ronald Fisher (1890–1962), who developed maximum likelihood estimation (MLE) to infer parameters of probability distributions. Fisher's work allowed statisticians to estimate relationships between variables and quantify uncertainty.

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

are not fully understood. For example, in modern deep learning, neural networks can recognize faces with high accuracy but offer little interpretability in how they make decisions. Unlike inference, which aims to understand why a pattern exists, prediction is about making the best possible guess given the available data. A key contributor to predictive modeling was Andrey Kolmogorov (1903–1987), who formalized probability theory and stochastic processes. His work laid the foundation for modern probabilistic models used in machine learning and artificial intelligence. Later, Vapnik and Chervonenkis (1971) developed statistical learning theory, which introduced VC dimension, a measure of a model's ability to generalize. This work led to the development of Support Vector Machines (SVMs), a key predictive tool in machine learning.

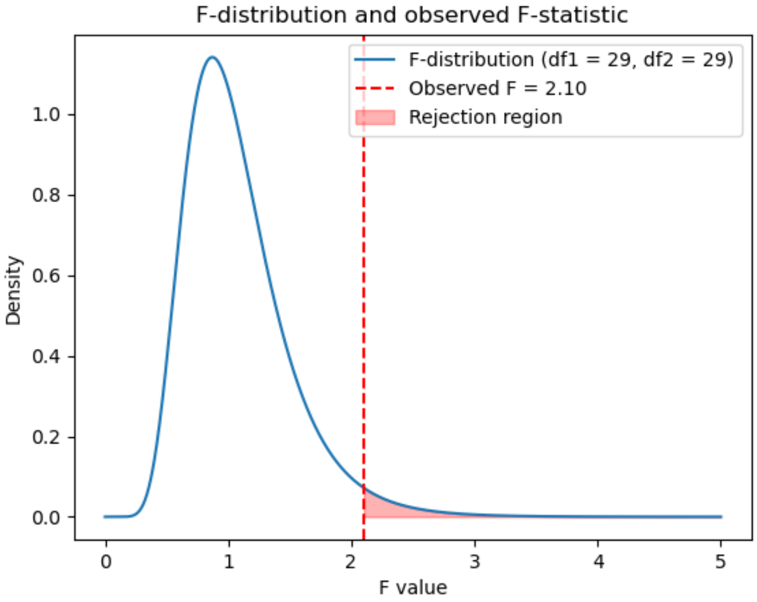

Figure 4.8: Representation of the Fisher distribution for given F_{obs} value.

While Breiman's paper emphasized the divide between the two approaches, modern data science integrates both. In many cases, understanding causal relationships (inference) can improve predictive performance, and accurate prediction can guide further investigations into causality. For example, epidemiologists use inference to understand how diseases spread, but they also use predictive models to forecast outbreaks. In finance, analysts infer factors affecting stock prices while using machine learning to predict market trends. By combining both approaches, we gain a more comprehensive understanding of data, allowing for both insightful interpretations and powerful predictions.

4.3.5 Compare distributions - χ^2 test

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

$$f(x; d_1, d_2) = \frac{s_1}{s_2},$$

where $\bar{x}$ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size.

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

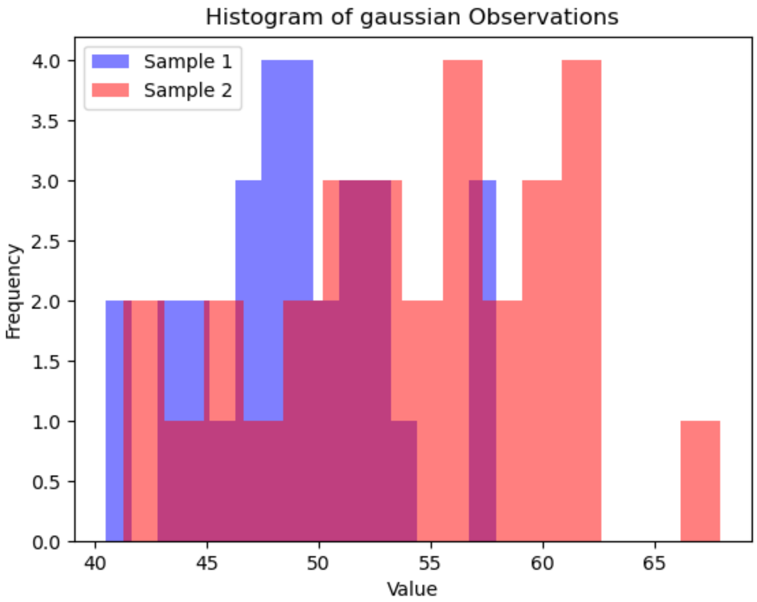

Figure 4.9: Observations following gaussian distribution.

Inference seeks to explain why and how variables relate. The key idea is causality and interpretability: given a dataset, inference aims to answer questions such as: Does smoking cause lung cancer, or is the correlation due to other confounding factors? How does an increase in temperature affect ice cream sales? What are the most significant predictors of house prices? Statistical techniques such as hypothesis testing, confidence intervals, and regression analysis help researchers understand these relationships. The emphasis is on estimating parameters and determining statistical significance rather than simply making accurate predictions. A classic example is Sir Ronald Fisher (1890–1962), who developed maximum likelihood estimation (MLE) to infer parameters of probability distributions. Fisher's work allowed statisticians to estimate relationships between variables and quantify uncertainty.

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

are not fully understood. For example, in modern deep learning, neural networks can recognize faces with high accuracy but offer little interpretability in how they make decisions. Unlike inference, which aims to understand why a pattern exists, prediction is about making the best possible guess given the available data. A key contributor to predictive modeling was Andrey Kolmogorov (1903–1987), who formalized probability theory and stochastic processes. His work laid the foundation for modern probabilistic models used in machine learning and artificial intelligence. Later, Vapnik and Chervonenkis (1971) developed statistical learning theory, which introduced VC dimension, a measure of a model's ability to generalize. This work led to the development of Support Vector Machines (SVMs), a key predictive tool in machine learning.

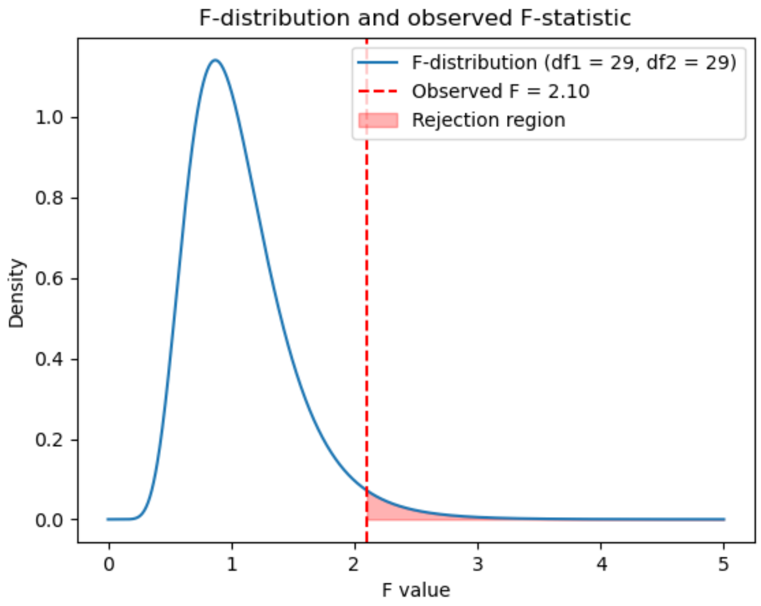

Figure 4.10: Representation of the Fisher distribution for given F_{obs} value.

While Breiman's paper emphasized the divide between the two approaches, modern data science integrates both. In many cases, understanding causal relationships (inference) can improve predictive performance, and accurate prediction can guide further investigations into causality. For example, epidemiologists use inference to understand how diseases spread, but they also use predictive models to forecast outbreaks. In finance, analysts infer factors affecting stock prices while using machine learning to predict market trends. By combining both approaches, we gain a more comprehensive understanding of data, allowing for both insightful interpretations and powerful predictions.

4.4 Parametric and non-parametric

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

Inference seeks to explain why and how variables relate. The key idea is causality and interpretability: given a dataset, inference aims to answer questions such as: Does smoking cause lung cancer, or is the correlation due to other confounding factors? How does an increase in temperature affect ice cream sales? What are the most significant predictors of house prices? Statistical techniques such as hypothesis testing, confidence intervals, and regression analysis help researchers understand these relationships. The emphasis is on estimating parameters and determining statistical significance rather than simply making accurate predictions. A classic example is Sir Ronald Fisher (1890–1962), who developed maximum likelihood estimation (MLE) to infer parameters of probability distributions. Fisher's work allowed statisticians to estimate relationships between variables and quantify uncertainty.

Prediction focuses on accuracy and generalization rather than explaining causality. The goal is to create a model that performs well on new, unseen data, even if the underlying relationships between variables are not fully understood. For example, in modern deep learning, neural networks can recognize faces with high accuracy but offer little interpretability in how they make decisions. Unlike inference, which aims to understand why a pattern exists, prediction is about making the best possible guess given the available data. A key contributor to predictive modeling was Andrey Kolmogorov (1903–1987), who formalized probability theory and stochastic processes. His work laid the foundation for modern probabilistic models used in machine learning and artificial intelligence. Later, Vapnik and Chervonenkis (1971) developed statistical learning theory, which introduced VC dimension, a measure of a model's ability to generalize. This work led to the development of Support Vector Machines (SVMs), a key predictive tool in machine learning.

While Breiman's paper emphasized the divide between the two approaches, modern data science integrates both. In many cases, understanding causal relationships (inference) can improve predictive performance, and accurate prediction can guide further investigations into causality. For example, epidemiologists use inference to understand how diseases spread, but they also use predictive models to forecast outbreaks. In finance, analysts infer factors affecting stock prices while using machine learning to predict market trends. By combining both approaches, we gain a more comprehensive understanding of data, allowing for both insightful interpretations and powerful predictions.

4.5 Comparing data and normalization

Comparing data and normalization

Chapter 5

Introduction to bayesian statistics

5.1 The Bayes' theorem

The Bayes' theorem.

5.2 Bayesian vs frequentist

Bayesian vs frequentist.

5.3 Bayesian statistics

Bayesian statistics.

Chapter 6

Stochasticity and Markov processes

6.1 Stochasticity and Markov processes

6.2 Markov chains and hidden Markov models

6.3 Modern applications

Bibliography

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

[2] Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics* (4th ed.). Pearson, 2012.

[3] J. A. F. McFadden. *The Philosophy of Statistics*. Wiley-Blackwell, 2011.