Imperial College London

Introduction to probability theory and statistical inference

Jesús Urtasun Elizari

Research Computing and Data Science

May 8, 2025

Contents

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 deep mathematical training.

First, we will introduce the idea of probability and random events with simple and intuitive examples, and we will see how different approaches have been used to model information and chance. Then we will discuss a series of mathematical ways to formally define random processes, also referred to as *stochastic*. We will introduce the idea of *distribution*, uncertainty and variability, and we will learn how to build quantities - we will call them *estimators*, or *expected values* - that represent the information we have about such random measurements.

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, and further stochasticity with the so-called Markov processes.

Ideas such as stochasticity and randomness have deep historical roots. Ancient civilizations, from Babylonians, Egyptians, and Greeks, faced uncertainty in games of chance, commerce, and divination, among others. 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 mathematicians like Gerolamo Cardano (1501–1576) made early contributions by analyzing gambling problems and laying the groundwork for probability theory.

Such intuitions properly formalized probability as a mathematical discipline 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. All these names and topics will be covered in detail through chapters 2 and 3.

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 branches of mathematics that aim to quantify uncertainty, chance and information in the random events, *much earlier* than computers, coding languages, Python, R or P-values were even conceived. The data-oriented, practical ways in which we nowadays focus probability and stats, heavily relying on computation, is just a consequence of the fact that automatized measurements are nowadays available and trendy in modern times.

As one could guess already, a proper understanding of such topics - uncertainty, variation, probability, chance - can be applied to describing a vas amount of real-world phenomena, ranging from gambling and statistical inference, to data analysis in physics, biology, 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 this with a bit more detail in Chapter 1.

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 defining a couple of quantities most people are already familiar with, and for which they may have some intuition, as a warmup example. 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 - let's call it x - that can be anything (position at a given time, energy of some system, concentration, ...). We repeat the measurement three times and we get first 1, then 2, and the last time 3. That will be our set of observations, or our *sample*, x_1. We could write it as a list, or a *vector* - in the following way as

$$x_1 = \{1, 2, 3\}.$$

Keep in mind that from the mathematics perspective the word *vector* has a 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 define the *mean* - or *average* - $\bar{x}$ of an arbitrary large sample of N observations, as the sum of all elements divided by the total.

$$\bar{x} = \frac{1}{N}(x_1 + x_2 + ... + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.1}$$

Here 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 x_1, which has just $N = 3$ observations, we get

$$\bar{x}_1 = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{1}{3}(1 + 2 + 3) = 2$$

As we see, the mean is just a quantity that captures some information about the "central" value, where the bulk of event are. In a similar way, we can define the *variance* as a quantity that captures how far are the elements of the sample from the mean value.

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2. \tag{1.2}$$

Again, just by substituting that expression for our set x_1, which has just $N = 3$ observations, we get

$$s_1^2 = \frac{1}{3-1}\sum_{i=1}^{3}(x_i - \bar{x})^2 = \frac{1}{2}\big((1-2)^2 + (2-2)^2 + (2-3)^2\big) = \frac{1}{2}(1+0+1) = 1$$

which we could interpret as, on average, the elements of the list being *one unit* away from the mean. As an exercise, try to compute both the mean and variance for a second sample, let's say

$$x_2 = \{4, 5, 6\}.$$

Buy substituting in the general expressions of $\bar{x}$ and s^2 you should get the following results:

$$\bar{x}_2 = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{1}{3}(4+5+6) = 5$$

$$s_2^2 = \frac{1}{3-1}\sum_{i=1}^{3}(x_i - \bar{x})^2 = \frac{1}{2}\big((4-5)^2 + (5-5)^2 + (6-5)^2\big) = \frac{1}{2}(1+0+1) = 1$$

Again, our mean $\bar{x}_2 = 5$ encodes the information about the "central" value, where the bulk of event are. The variance $s_2^2 = 1$ indicates that, as in the previous example, .
Another useful quantity used to characterize variability is the so called *standard deviation*, which is just the square root of the variance

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{1.3}$$

Sometimes it is useful to use the standard deviation and sometimes the the variance, depending on the question and topic [...]

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* [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].

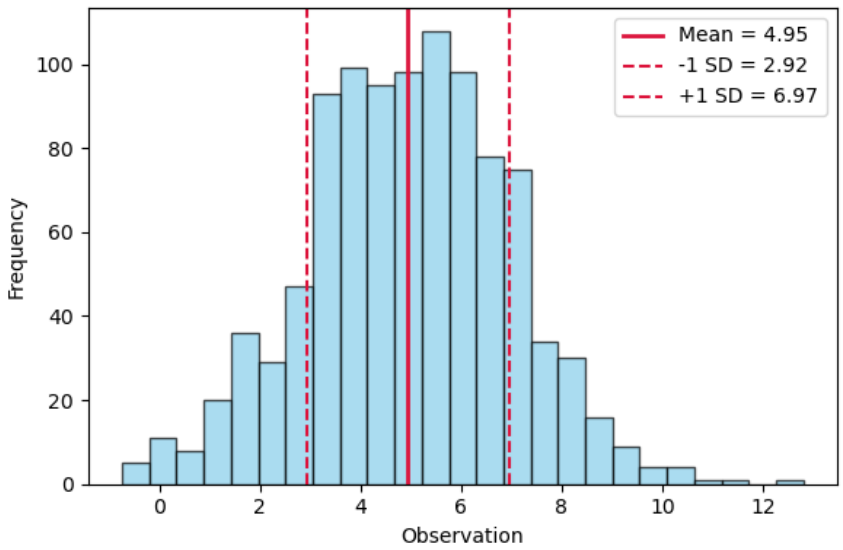

Figure 1.1: Histogram representing the mean and standard deviation for a set of gaussian observations. The mean shows the central value where the bulk of events lie, where the standard deviation is a measure of the variability, how spread the observations are with respect to the mean

Chapter 2

Probability and random events

2.1 What is probability?

As already mentioned in the introduction, probability is a branch of mathematics dealing with information and random events. Hence, a fair question to begin with, would be what *are* random events? Random events, also referred to as *stochastic*, we will mean simply a process whose output we *ignore*. As classic 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.

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

The example of the coin, where we have just two possible results, is what we will call a *Bernoulli* trial, which we will describe soon, but let's use it as a prior example now to introduce the idea of probability.

The cases in which I am certain, of either one case of the other, are clear. But for the third one, where we assign a value to the probability which is not 0 or 1, we should stop for a second. When we say that the probability of getting heads - or tails - in a normal coin that is not biased is $P = \frac{1}{2}$, 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 (e.g., heads), and the other half the remaining result (e.g., tails). 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 n, and the number of total trials N.

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

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

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

Ideally we expect that these frequencies, as we increase the number of repetitions, would approach a perfect $\frac{1}{2}$. 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 recover the total. That means, at least one of the scenarios will happen.

The formal definition of unitarity can be written as follows. Let's denote all possible outcomes of an experiment x_1, x_2, ..., x_n. In the case of coins these will be just $x_1 = H$, $x_2 = T$, and with dice, $x_1 = 1$, $x_2 = 2$, ..., $x_6 = 6$. By *unitarity*, we mean that the sum of probabilities of all possible outcomes add up to 1.

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

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.

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 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 - even better, *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*.

2.2.1 Bernoulli distribution

The simplest case we can think of is the **Bernoulli trial**, named after Swiss mathematician Jacob Bernoulli in late 1600s. A Bernoulli trial is a random experiment with exactly two possible outcomes: *success*, usually labeled as 1, and *failure*, labeled as 0. The probability of success is denoted by p, and the probability of failure is $1 - p$. Mathematically, for a single Bernoulli trial with random variable x,

$$P(x = 1) = p \quad \text{and} \quad P(x = 0) = 1 - p, \tag{2.3}$$

where $p \in [0, 1]$. Note that both probabilities do sum 1, and hence if they properly obey the unitarity property. As well, you can see that this is a generalization of the case of the coin, in which the two outcomes had the same probability $p = 0.1$.

Jacob Bernoulli (1655–1705) was one indeed of the pioneers of probability theory. His work *Ars Conjectandi*, published posthumously in 1713, laid the groundwork for the law of large numbers and formalized many concepts still used today.

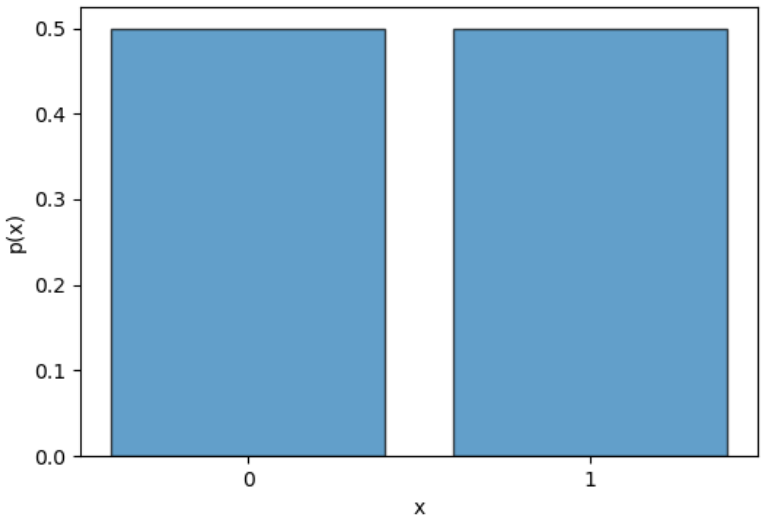

Figure 2.1: Representation of the bernoulli distribution of a random variable x, given the total number of trials n and the individual probability of success p.

Example 1: Coin Toss

A fair coin toss is a Bernoulli trial with

$$p = P(\text{Heads}) = P(\text{Tails}) = 0.5. \tag{2.4}$$

And we can model it as:

$$x = \begin{cases} 1 & \text{if Heads} \\ 0 & \text{if Tails} \end{cases} \tag{2.5}$$

Bernoulli trials form the basis for more complex models such as the **Binomial distribution**, which models the number of successes in a fixed number of independent Bernoulli trials.

2.2.2 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:

$$P(x; n, k) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{2.6}$$

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.

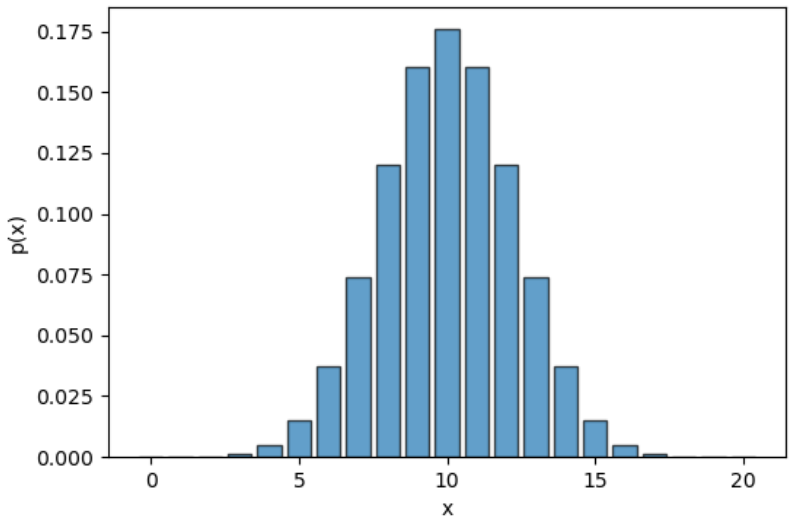

Figure 2.2: Representation of the binomial distribution of a random variable x, given the total number of trials n and the individual probability of success p.

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

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

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 ($p = 0.5$):

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

2.2.3 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:

$$P(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \tag{2.7}$$

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

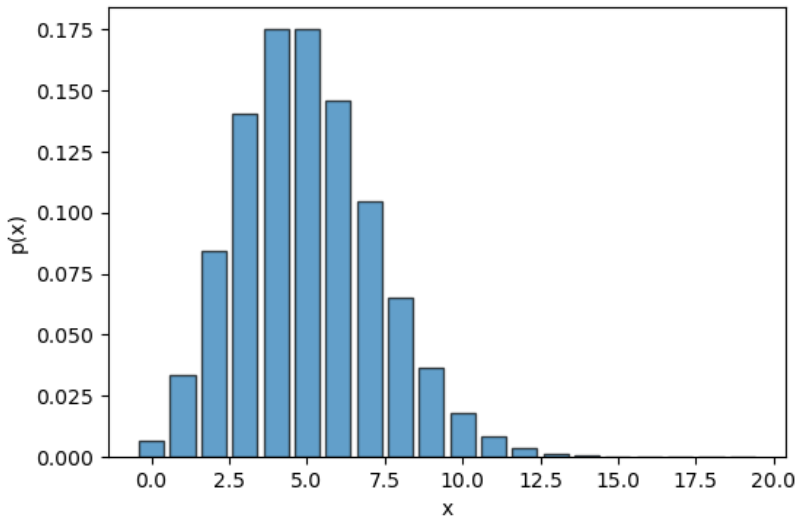

Figure 2.3: Representation of the Poisson distribution of a random variable x, given the number of observations λ as a parameter.

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.

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

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.

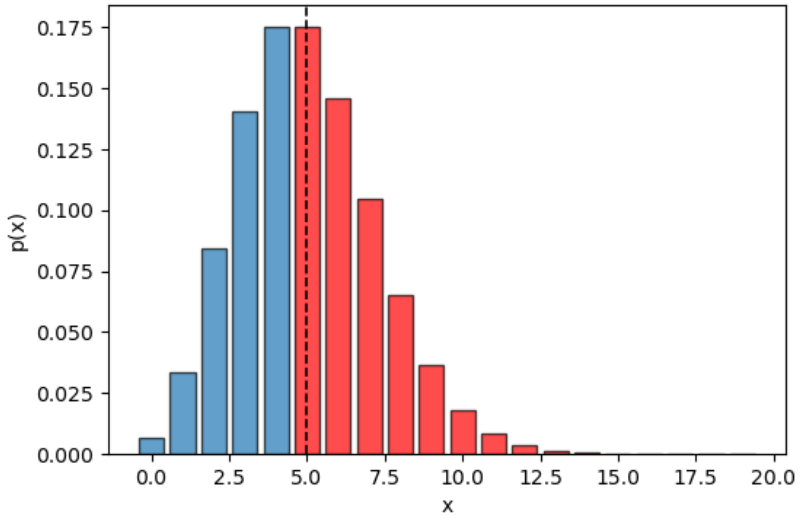

Figure 2.4: Representation of the Poisson distribution of a random variable x, given the number of observations λ as a parameter.

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

$$P(x \leq 5; \lambda = 3) = P(x = 0; \lambda = 3) + P(x = 1; \lambda = 3) + P(x = 2; \lambda = 3) + P(x = 3; \lambda = 3) + P(x = 4; \lambda = 3) + P(x = 5; \lambda = 3)$$

Given unitarity, we can just compute it as

$$P(x > 5; \lambda = 3) = 1 - P(x \leq 5; \lambda = 3) = 1 - CDF(5; \lambda = 3).$$

2.2.4 Uniform distribution

The uniform distribution represents a scenario where all outcomes in an interval $[a, b]$ are equally likely. The probability of observing a particular result x in a given range $[a, b]$ is:

$$f(x; a, b) = \frac{1}{b-a}, \quad a \leq x \leq b. \tag{2.8}$$

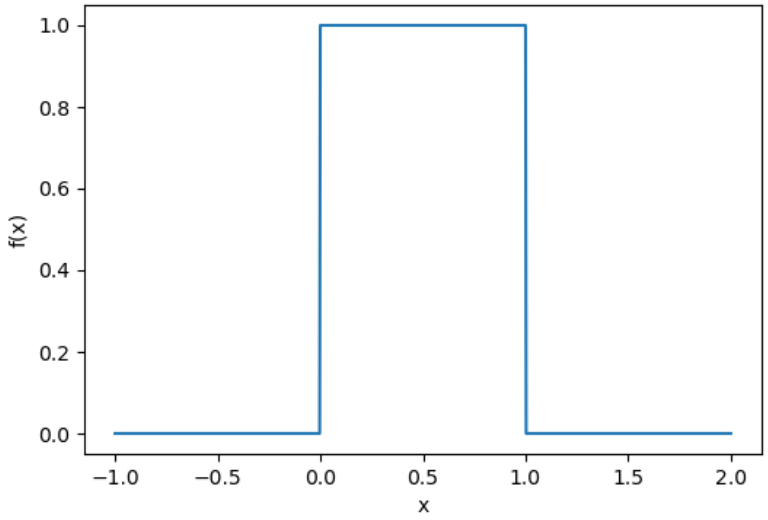

Figure 2.5: Representation of the uniform distribution of a random variable x, given the boundary parameters a, b.

2.2.5 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.

The probability density function is:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.9}$$

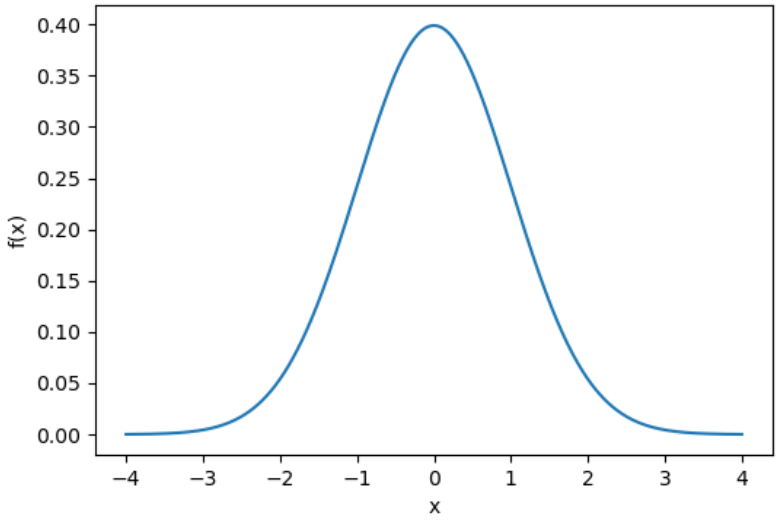

Figure 2.6: Representation of the gaussian distribution of a random variable x, given the mean value μ and standard deviation σ parameters.

2.2.6 Exponential distribution

The exponential distribution models waiting times between. 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:

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0. \tag{2.10}$$

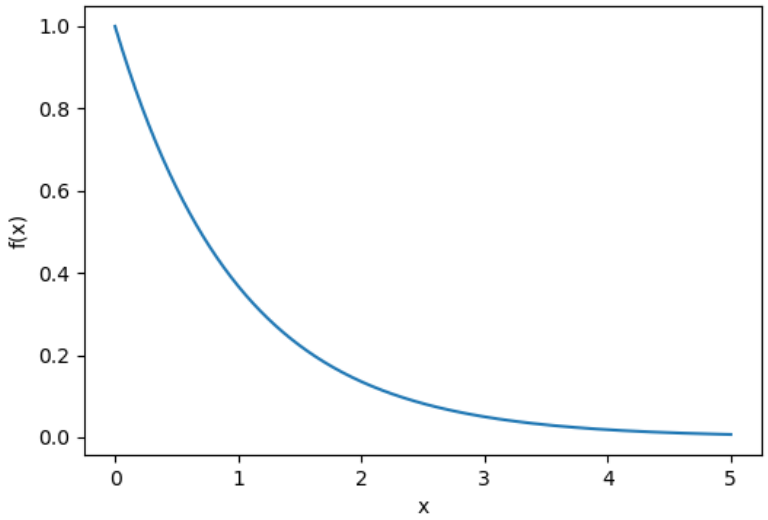

Figure 2.7: Representation of the exponential distribution of a random variable x, given the decay rate λ.

2.3 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*. [...]

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:

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

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:

$$P(x = x_0) = \frac{\text{number of times I get } x_0}{\text{number of times I get any other result}} = \frac{n}{\infty} = 0 \tag{2.12}$$

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.

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 - prediction - 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 a given set of assumptions are compatible with the obtained 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 some set of observations, 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?

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.

Statistical techniques such as hypothesis testing, confidence intervals, and regression analysis aim to understand and describe these relationships. The emphasis lies 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.

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.

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 The Law of Large Numbers

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, such as 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.

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:

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

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.

- **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.,

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

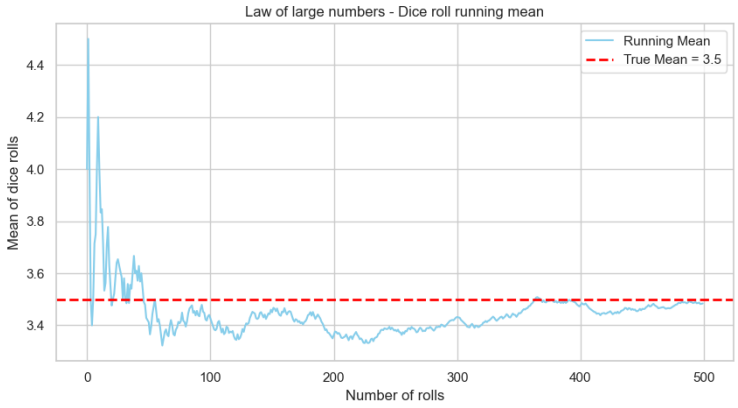

Figure 3.1: Representation of the law or large numbers. The sample mean tends to the population mean as the number of rolls n increases.

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 The Central Limit Theorem

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.

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 $n \to \infty$.

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.

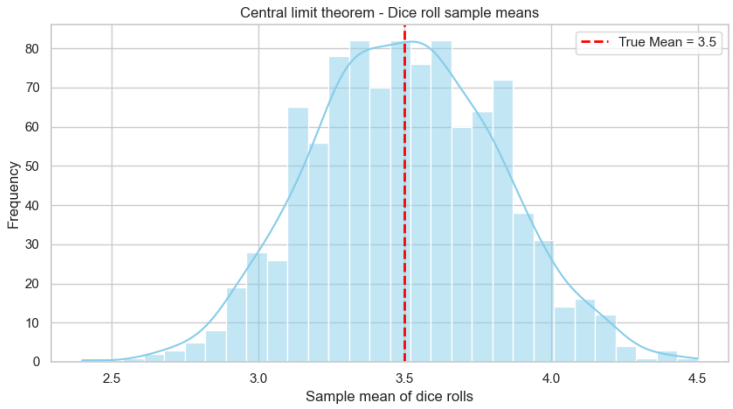

Figure 3.2: Representation of the law or large numbers. The sample mean follows a gaussian distribution as the sample size n increases.

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 $\mu = 3.5$.

- 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.

3.5 Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) is a cornerstone of modern statistical inference. Developed in the early 20th century by Sir Ronald A. Fisher, MLE provides a systematic framework for estimating the parameters of a probabilistic model. Fisher introduced the method in the 1920s, formalizing it as a rigorous alternative to the method of moments and laying the groundwork for much of classical and modern statistical theory. MLE has since become one of the most widely used estimation techniques due to its generality, mathematical tractability, and strong theoretical properties. It applies to a broad class of models, including both discrete and continuous distributions, and serves as the basis for many advanced statistical methods, including Generalized Linear Models (GLMs), Bayesian inference (as the likelihood term), and machine learning algorithms.

3.5.1 1. Motivation and Intuition

MLE seeks the parameter θ that makes the observed data most probable under the assumed model. In other words, it chooses the parameter that maximizes the likelihood function:

$$L(\theta) = P(X_1 = x_1, \ldots, X_n = x_n \mid \theta)$$

Example: For a Bernoulli distribution with unknown probability p, the likelihood of observing a sequence of 0s and 1s is:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

3.5.2 2. The Likelihood and Log-Likelihood Functions

For independent and identically distributed data $X_1, \ldots, X_n$ with density or mass function $f(x;\theta)$, the likelihood function is:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

To simplify differentiation, we often use the **log-likelihood**:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

3.5.3 3. Finding the MLE

To find the MLE $\hat{\theta}$:

1. Write down the log-likelihood $\ell(\theta)$.
2. Take the derivative with respect to θ: $\frac{d\ell}{d\theta}$.
3. Solve $\frac{d\ell}{d\theta} = 0$ to find critical points.
4. Check which value maximizes the likelihood (often via the second derivative or boundary checks).

Example: Bernoulli MLE
For $X_i \sim \text{Bernoulli}(p)$,

$$\ell(p) = \sum_{i=1}^{n} [x_i \log(p) + (1-x_i)\log(1-p)]$$

Taking derivative:

$$\frac{d\ell}{dp} = \sum_{i=1}^{n} \left[\frac{x_i}{p} - \frac{1 - x_i}{1 - p} \right] = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1 - p}$$

Solving yields:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

3.5.4 4. Properties of the MLE

The importance of MLE lies not only in its general applicability but also in its powerful theoretical properties. Under regularity conditions, MLEs are asymptotically optimal estimators in the sense that they:

- **Consistency:** $\hat{\theta} \to \theta$ as $n \to \infty$
- **Asymptotic Normality:** $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$, where $I(\theta)$ is the Fisher information
- **Efficiency:** Asymptotically achieves the Cram'er-Rao lower bound
- **Invariance:** If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$ for any differentiable function g

These properties make MLE a preferred method in both theoretical and applied statistics, especially for large-sample inference. In practice, these properties justify the use of MLE even when exact finite-sample distributions are hard to derive.

3.5.5 5. Application to Generalized Linear Models (GLMs)

Generalized Linear Models (GLMs) are an important class of models that extend linear regression to non-normal response variables by using a link function and a distribution from the exponential family. MLE plays a central role in fitting GLMs because the estimation of the model parameters is achieved by maximizing the likelihood of the observed responses. For instance, in logistic regression—used for binary outcomes—the log-odds of success is modeled as a linear combination of predictors:

Example: Logistic Regression

For binary response data, logistic regression models the log-odds as a linear function of predictors:

$$\log\left(\frac{p}{1 - p}\right) = \beta_0 + \beta_1 x$$

MLE is used to estimate the coefficients β_0, β_1 by maximizing the binomial log-likelihood.

Exercises

1. **Basic Derivation:** Show that the MLE for the rate parameter λ of an exponential distribution $f(x; \lambda) = \lambda e^{-\lambda x}$ is $\hat{\lambda} = 1/\bar{x}$.

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 - first - and experiment - after - 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

The term *hypothesis testing* is usually used to refer a broad set of tools addressing parameter estimation, inference, and various exploratory analysis on random measurements and observations. It was first coined by British mathematicians Pearson and Fisher [...] in early XX^{th} century. In the last decades, hypothesis testing, hypothesis test, statistical inference - sometimes referred to as exploratory analysis - has gained popularity and become one of the standards in most experimental sciences, given the automatization of experiments and the large amounts of data available.

Generally speaking, we will use hypothesis testing as a set of rules, almost as an recipe or an algorithm, to face and interpret data. Once we have covered the idea of parameter estimation, sample distributions, and the idea of estimators, we will now formulate hypothesis on the true - *unkwnown* - parameters, and then build *statistic tests* to quantify how far - or close - are these hypothesized values from the observed - experimental, sample - values. And finally, we will quantify how certain we are about the values obtained - how *significant* they are - computing the *p-value*, standing from Pearson value.

The general approach we will follow, regardless of the kind of question we are after and the observations made, can be summarized as follows:

- Formulate *null* hypothesis H_0 and *alternative* hypothesis H_1 about the *true* - *population* parameters, generally for the mean or variance, *prior to experiment*.
- Collect data, make observations, make measurements.
- Compute *informative quantities* from our observed values, normally referred to as *statistics*, or *statistic tests*
- Compute p-value, probability of *given the null hypothesis was true* obtained a value at leas as extreme as the one we obtained for our statistic test.

- Accept or reject the null hypothesis, based on the p-value.

About statistc tests [...].

About p-values [...]

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

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 was developed by William Gosset, a statistician working at the Guinness factory, in 1908 [1]. Due to restrictions due to the agreements with the company, he was not able to submit it with is name to the scientific journal [...], and hence it was published under the pseudonym *student's t* algorithm. Originally developed to quantify if certain concentrations of barley / rye [...].

The student's t is used to compare the sample mean $\bar{x}$ to ha hypothesized value μ. It assumes that the sample data are drawn from a normally distributed population, hence it is an example of a *parametric* test. We will discuss more about parametric and non-parametric observations, and how to test for normality further in the chapter. The test statistic is given by:

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

where $\bar{x}$ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size. It is built in such a way that, as the sample mean $\bar{x}$ gets closer to the hypothesized value μ, the t-variable approaches zero.

Then, given some data was observed and and we obtained a specific value for our t - let's call it $t\ obs$, to compute a p-value we just need to compute what was the probability of that particular value. To do that, we just recall our t variable was indeed a random variable depending on our random observations, which produced some random sample mean and variance, and some degrees of freedom $n-1$

$$p = P\left(t > tobs\right) = 2 \cdot \int_{|t|}^{\infty} f_{T_{n-1}}(x)\, dx = 2 \cdot \left[1 - F_{T_{n-1}}(|t|)\right]$$

Being $f_{T_{n-1}}$ the PDF of the t variable, the *Student's t distribution* with $n-1$ degrees of freedom, and $F_{T_{n-1}}$ the corresponding cumulative distribution, as we discussed in chapter 2 [...]. Here, we are computing the probability of t being greater than the one we obtained, and we do that just by integrating the t-distribution [...]. Note that here we are computing a 2-sided p-value, hence te factor 2 at the beginning.

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.

The computation of the p-value:

$$p = P\left(t > tobs\right) = 2 \cdot \int_{|t|}^{\infty} f_{T_{df}}(x)\, dx = 2 \cdot \left[1 - F_{T_{df}}(|t|)\right]$$

Being $f_{T_{n-1}}$ the PDF of the t variable, the *Student's t distribution* with $n_1 + n_2 - 1$ degrees of freedom, and $F_{T_{n-1}}$ the corresponding cumulative distribution Note here we assume equal variances. For Welch's t-test (unequal variances), use the same form, but with Welch-adjusted [...]

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 F statistic is a ratio of two independent variance estimates, each scaled by their respective degrees of freedom. It is used to test whether group variances (or group means, in ANOVA) differ significantly. The general form of the F statistic is:

$$F = \frac{S_1^2/\nu_1}{S_2^2/\nu_2}$$

where s_1^2 and s_2^2 are the sample variances, and the degrees of freedom are $d_1 = n_1 - 1$ and $d_2 = n_2 - 1$.

The computation of the p-value:

$$p = \sum_{\text{all tables as or more extreme}} \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

Under the null hypothesis, the F statistic follows the F-distribution:

$$F \sim F(\nu_1, \nu_2)$$

and the p-value is computed as the upper-tail probability:

$$p = P(F_{\nu_1,\nu_2} \geq F_{\text{obs}}) = \int_{F_{\text{obs}}}^{\infty} f_{F_{\nu_1,\nu_2}}(x)\, dx$$

Not to bo

4.3.4 Compare variation on more than two groups - ANOVA

The so-called Analysis of Variance, one way ANOVA, or just ANOVA, is used to determine whether the variation of a dataset comes primary from variation within the samples themselves, or from variation between the groups. It is an extension of the Fisher test, where the F statistic is computed as:

$$f(x; d_1, d_2) = \frac{s^2_{\text{between}}}{s^2_{\text{within}}},$$

where s_1^2 and s_2^2 are the sample variances, and the degrees of freedom are $d_1 = n_1 - 1$ and $d_2 = n_2 - 1$.

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{SS_{\text{between}}/(k-1)}{SS_{\text{within}}/(N-k)}$$

where:

- SS_{between} is the sum of squares between groups,
- SS_{within} is the sum of squares within groups,
- k is the number of groups,
- N is the total number of observations.

The computation of the p-value:

$$p = \int_F^\infty f_{F_{df_1, df_2}}(x)\, dx = 1 - F_{F_{df_1, df_2}}(F)$$

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

The so-called Pearson χ^2-test is used to determine whether the set of observations is significantly different from some expected - or hypothesized - values. The χ^2 statistic is given by:

$$\chi^2(x; N-1) = \sum_{i=1}^{N} \frac{(O_i - Ei)^2}{E_i},$$

It is also used to test for normality and evaluate the goodness of a fit [...].

The computation of the p-value:

$$p = \int_{\chi^2}^{\infty} f_{\chi^2_{df}}(x)\, dx = 1 - F_{\chi^2_{df}}(\chi^2)$$

4.4 Parametric and non-parametric

Parametric and non-parametric

4.5 Comparing data and normalization

Comparing data and normalization

Chapter 5

Linear models and GLMs

Chapter 6

Introduction to bayesian statistics

6.1 The Bayes' theorem

The Bayes' theorem.

6.2 Bayesian vs frequentist

Bayesian vs frequentist.

6.3 Bayesian statistics

Bayesian statistics.

Chapter 7

Stochasticity and Markov processes

7.1 Stochasticity and Markov processes

Stochasticity and Markov processes

7.2 Markov chains

Markov chains

7.3 Hidden Markov models

Hidden Markov models

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.