An overview of Probability, Statistics and Stochastic Processes

D. Pinheiro*

CEMAPRE, ISEG
Universidade Técnica de Lisboa
Rua do Quelhas 6, 1200-781 Lisboa
Portugal

May 18, 2012

Abstract

The present manuscript constitutes the lecture notes for the "Probability, Statistics and Stochastic Processes" module of the PhD program on Complexity Science, held at ISCTE-IUL during April and May 2012. The aim of the notes is to provide some auxiliary material for the students to follow this 10 hour length module, devoted to the study of the Probability theory and Stochastic Processes, as well as Statistics. A good working knowledge of calculus in several variables and linear algebra is desirable.

Contents

*Email: dpinheiro@iseg.utl.pt

1 Introduction

In order to pursue a deeper understanding of complex behaviour in science and social science, one should be prepared to accept the fact that most real world phenomena, if not all, have some degree of uncertainty attached to it. The emergence of uncertainty may be due to a number of reasons. One possibility is that the object of study has a random nature by itself. Some simple examples of this kind of phenomena include hazard games, quality control, and particle detection. On the other hand, even deterministic experiments may exhibit some sort of uncertainty which, for instance, may be due to the presence of measurement errors or some fluctuation of environmental factors. For a typical experiment, there may even be several distinct

sources of uncertainty affecting the corresponding outcome. It is then relevant to be able to identify and describe the influence of such random phenomena in the outcomes of the experiment under consideration. The mathematical tools to address such problems are part of Probability theory and Statistics.

Probability theory is the branch of mathematics devoted to the study of random phenomena. Its initial development was motivated by the analysis of hazard games, dating back to the XVI and XVII centuries. Prominent figures responsible for this initial development include Cardano, Fermat and Pascal. The current form of the theory is mainly due to Kolmogorov, responsible for the axiomatization of the theory in the early XX century. A stochastic process is one of the objects of study in Probability theory, used to study the evolution with time of some random phenomena.

Statistics is one of the key applications of Probability theory, providing tools for the quantitative analysis of large sets of data. Its main goal is to study the appropriate procedures for the collection, organization, analysis, and interpretation of data.

An effort has been made to make this short course self contained. The aim is to provide the student with an overview of the subject, describing some of the main concepts and results of the theory, complemented by some illustrative examples. In order to cover a wider range of subjects, proofs and technical details are avoided in the present text, but can be found in the references provided at the end of this notes.

2 Probability

This section is devoted to the introduction of some of the main concepts in Probability Theory. We start by describing its current formulation in terms of probability spaces, and then we move on to discuss concepts such as random variables and some special probability distributions. We finish the section with two key results: the law of large numbers and the central limit theorem.

2.1 Probability spaces

Probability theory is the branch of mathematics concerned with the study of random experiments. Such experiments share the property that their outcomes is not predetermined, i.e. if a random experiment is repeated under exactly the same conditions, the resulting outcome is not necessarily equal. As trivial examples one may think of tossing a coin, rolling a dice, or drawing a card from a 52-card deck.

The first object to be introduced in probability theory is the space of elementary outcomes, which we will call *sample space* and will denote by Ω from now onwards. The sample space Ω is a non-empty set, whose elements $\omega \in \Omega$ are called *elementary outcomes* or *elementary events*.

Example Here are several simple examples of random experiments with the corresponding sample spaces.

(i) Consider an experiment involving a single coin toss. There are two possible outcomes, heads (H) and tails (T). The sample space is $\Omega = \{H, T\}$.

(ii) Consider another experiment involving two coin tosses. The outcome will be a string with two elements, each representing either heads or tails. The sample space is

$$\Omega = \{HH, HT, TH, TT\} \ .$$

(iii) If we consider the experiment of rolling a single dice, the elementary outcomes are "face 1", "face 2", up until "face 6" and the sample space is

$$\Omega = \{\text{“face 1”}, \text{“face 2”}, \ldots, \text{“face 6”}\} \ .$$

(iv) For the experiment of rolling a dice until we obtain "face 6", the elementary outcomes are "1" (to obtain "face 6" at the first time the dice is rolled), "2" (to obtain "face 6" at the second time the dice is rolled), and so on up until "∞" (representing the outcome "face 6" is never obtained). In this case the sample space is given by

$$\Omega = \{\text{“1”}, \text{“2”}, \text{“3”}, \ldots, \text{“}\infty\text{”}\} \ .$$

(v) Another experiment may be to measure the height of a randomly chosen 10 year old Portuguese child. The elementary outcomes of such experiment are positive real numbers corresponding to the child height. The sample space is $\Omega = \mathbb{R}^+$, the set of positive real numbers. One could also think of measuring the height and weight of a randomly chosen 10 year old Portuguese child. In this case, the elementary outcomes are pairs of positive real numbers corresponding to the child height and weight and the sample space is $\Omega = \mathbb{R}^+ \times \mathbb{R}^+$.

To complete this set of examples, it should be remarked that examples (i), (ii) and (iii) have finite sample spaces, (iv) has a discrete (countable) infinite sample space, and (v) has a continuous (non-countable) sample space.

The next object that needs to be introduced for the study of probability theory is the notion of event. A naive definition would be to say that an event is a subset of the sample space Ω. However, some additional structure is needed to ensure that set operations such as union, intersection and complementary are closed. Such structure is provided by the notion of σ-algebra.

Before moving on to define σ-algebra, we recall the definition of the set operations mentioned above. Let A and B be subsets of Ω. The *intersection* of A and B, denoted by $A \cap B$, is the subset of Ω whose elements belong simultaneously to A and B, i.e. it is the set

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\} \ .$$

The *union* of A and B, denoted by $A \cup B$, is the subset of Ω whose elements belong to A or B, i.e. it is the set

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\} \ .$$

The *complementary* of A in Ω, denoted by A^c or $\Omega \backslash A$, is the subset of Ω whose elements do not belong to A, i.e. it is the set

$$A^c = \{\omega \in \Omega : \omega \notin A\} \ .$$

Similarly, one can define the *complementary* of A in B, denoted by $B\backslash A$, as the subset of Ω whose elements belong to B and do not belong to A, i.e. it is the set

$$B\backslash A = \{\omega \in B : \omega \notin A\} \ .$$

We are now ready to define σ-algebra.

Definition 2.1.1 (σ-algebra)**.** *A collection $\mathcal{F}$ of subsets of Ω is called a σ-algebra if the following three properties hold:*

1) $\Omega \in \mathcal{F}$*;*

2) $A \in \mathcal{F}$ *implies that* $A^c \in \mathcal{F}$*.*

3) $A_i \in \mathcal{F}$*,* $i \geq 1$*, implies that* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Elements of $\mathcal{F}$ are called measurable sets*, or* events*.*

A few more comments concerning properties of a σ-algebra $\mathcal{F}$. It follows easily from the previous definition that:

(i) the empty set $\emptyset$ is an element of $\mathcal{F}$;

(ii) if $A_1, \ldots, A_n \in \mathcal{F}$, then $\bigcap_{i=1}^{n} A_i \in \mathcal{F}$;

(iii) if $A, B \in \mathcal{F}$, then $B\backslash A \in \mathcal{F}$.

The simplest examples of a σ-algebra for a sample space Ω are the trivial σ-algebra $\underline{\mathcal{F}} = \{\emptyset, \Omega\}$ and the σ-algebra $\overline{\mathcal{F}}$ containing all the subsets of Ω. However, depending on the random experiment, other non-trivial σ-algebras may be considered.

Definition 2.1.2 (Measurable space)**.** *A measurable space is a pair $(\Omega, \mathcal{F})$, where Ω is a space of elementary outcomes and $\mathcal{F}$ is a σ-algebra of subsets of Ω.*

Having introduced the notion of a measurable space, we are now able to define what is meant by a probability measure. Shortly, a probability measure P in a measurable space $(\Omega, \mathcal{F})$ is a function that assigns real numbers in the interval $[0, 1]$ to events $A \in \mathcal{F}$. The rigorous definition is provided below.

Definition 2.1.3 (Probability measure)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space. A probability measure on $(\Omega, \mathcal{F})$ is a function $P : \mathcal{F} \to \mathbb{R}$ for which the following properties hold:*

(i) $P(\Omega) = 1$*;*

(ii) $P(A) \geq 0$ *for every* $A \in \mathcal{F}$*;*

(iii) for every sequence of events $\{C_i\}_{i \in \mathbb{N}} \subset \mathcal{F}$ *such that* $C_i \cap C_j = \emptyset$ *for* $i \neq j$ *we have*

$$P\left(\bigcup_{i=1}^{\infty} C_i\right) = \sum_{i=1}^{\infty} P(C_i) \ .$$

From the properties of a probability measure listed above, one can deduce several other well known properties. Let $A, B \in \mathcal{F}$. We have that:

- $P(\emptyset) = 0$;
- $P(A^c) = 1 - P(A)$;

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
- $P(B \backslash A) = P(B) - P(A \cap B)$;
- if $A \subseteq B$, then $P(A) \leq P(B)$.

Definition 2.1.4 (Probability space)**.** *A* probability space *is a triplet* $(\Omega, \mathcal{F}, P)$*, where* $(\Omega, \mathcal{F})$ *is a measurable space and* P *is a probability measure on* $(\Omega, \mathcal{F})$*. If* $C \in \mathcal{F}$*, the number* $P(C)$ *is called the* probability *of the event* C*.*

Example Recall the random experiments of the previous example.

(i) Single coin toss. The sample space is

$$\Omega = \{H, T\} \ .$$

We can endow Ω with the σ-algebra $\mathcal{F}$ defined by

$$\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$$

and a probability measure P assigning probabilities $P(\{H\}) = p$ and $P(\{T\}) = q$, with $p + q = 1$, to the elementary events of Ω. If the coin is balanced, then $p = q = 1/2$. Note that Ω admits only the two trivial choices for σ-algebra.

(ii) Two coin tosses. The sample space Ω given by

$$\Omega = \{HH, HT, TH, TT\}$$

can be endowed with the σ-algebra $\mathcal{F}_1$ defined by

$$\begin{aligned} \mathcal{F}_1 \ = \ & \Big\{\emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \\ & \{HT, TH\}, \{HT, TT\}, \{TH, TT\}, \{HH, HT, TH\}, \{HH, HT, TT\}, \\ & \{HH, TH, TT\}, \{HT, TH, TT\}, \Omega\Big\} \ . \end{aligned}$$

This is the largest possible σ-algebra for Ω. Other choices are possible and may be useful depending on the situation at hand. An alternative choice is

$$\mathcal{F}_2 = \Big\{\emptyset, \{HH, HT\}, \{TH, TT\}, \Omega\Big\} \ .$$

An easy exercise is to check that $\mathcal{F}_1$ and $\mathcal{F}_2$ are indeed σ-algebras of Ω and to think of two distinct situations modeled, respectively, by $\mathcal{F}_1$ and $\mathcal{F}_2$.
We will now define a probability measure for each one the σ-algebras above. We assume that the coin is balanced. For $\mathcal{F}_1$, define $P_1 : \mathcal{F}_1 \to [0, 1]$ by

$$P_1(\{HH\}) = P_1(\{HT\}) = P_1(\{TH\}) = P_1(\{TT\}) = \frac{1}{4} \ .$$

For $\mathcal{F}_2$, define $P_2 : \mathcal{F}_2 \to [0, 1]$ through

$$P_2(\{HH, HT\}) = P_2(\{TH, TT\}) = \frac{1}{2} \ .$$

Note that we have defined P_1 and P_2 by assigning probabilities to the smaller non-empty sets of $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. This is enough in this particular example. As an exercise, find the probabilities for the remaining events in $\mathcal{F}_1$. This should also provide a clue for the following two questions. Why could we proceed in this way in this example? Can you find an example for which this approach does not work?

(iii) Rolling a dice with six faces. This example is very similar to the one in item (ii). As an exercise, find at least two distinct σ-algebras and the corresponding probability measures under the assumption that the dice is properly balanced.

(iv) Rolling a dice until the outcome "face 6" is realized. We have seen above that the sample space is given by

$$\Omega = \{\text{“1”}, \text{“2”}, \text{“3”}, \ldots, \text{“}\infty\text{”}\} .$$

For a σ-algebra, one can take the set $\mathcal{F}$ of all subsets of Ω. Note that since Ω is a set with an infinite number of elements, so must be the case for $\mathcal{F}$. To endow the measurable space $(\Omega, \mathcal{F})$ with a probability measure, it is again enough to consider elementary events in $\mathcal{F}$. A good exercise is to check that the map given by

$$P(\{\text{“i”}\}) = \frac{1}{6}\left(\frac{5}{6}\right)^{i-1} \quad , \quad i \in \mathbb{N}$$

defines a probability measure on $(\Omega, \mathcal{F})$.

(v) The height of a randomly chosen 10 year old Portuguese child. As seen above, the sample space may be taken to be $\Omega = \mathbb{R}^+$. The most common choice for a σ-algebra on Ω is the Borel σ-algebra, the σ-algebra generated by the open sets in Ω. Its construction is slightly more subtle than the constructions of the previous examples, being outside the scope of this short course. General probability measures on Ω are no longer obtained by assigning probabilities to elementary events of Ω. However, some examples of probability measures (distributions) suitable to model this kind of random experiments will be discussed below.

A sample space Ω is said to be *discrete* if it has a finite or countable number of elements. One can always endow Ω with the σ-algebra $\mathcal{F}$ consisting of all the subsets of Ω. As in some of the previous examples, to define a probability measure on $(\Omega, \mathcal{F})$ it is enough to assign probabilities to its elementary events. However, such probabilities must satisfy a couple of consistency conditions:

(i) $P(\{\omega\} \geq 0$ for every $\omega \in \Omega$;

(ii) $\sum_{\omega \in \Omega} P(\{\omega\} = 1$.

We will now introduce two important concepts describing relations between two events and their probabilities.

Definition 2.1.5 (Independent events)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A, B \in \mathcal{F}$. The events A and B are* independent *if*

$$P(A \cap B) = P(A)P(B) .$$

From an heuristic point of view, two events are independent if the occurrence of one does not influence the probability of occurrence of the other. The trivial examples of independent events are the impossible event (empty set) and the certain event (full sample space), which are independent of every other event. For another very simple example, consider the random experiment consisting of two coin tosses. Clearly, the outcome of the first toss does not affect the outcome of the second toss. Thus, the event {the first toss outcome is tail} is independent of the

event {the second toss outcome is tail}. As an exercise, think of two or three more examples of independent events in distinct random experiments.

It should now be remarked that independency is a property of the probability measure, not just of events. For instance, two mutually exclusive (disjoint) events with positive probability are not independent. An easy exercise: why is this last statement true?

Definition 2.1.6 (Conditional probability)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A, B \in \mathcal{F}$ be such that $P(B) > 0$. The* conditional probability of A given B *is*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \ .$$

The concept of conditional probability provides the following information: $P(A|B)$ is the probability that an event A occurs based on the extra information that the event B also occurs. It enables one to reassess the probability of occurrence of an event when some additional information is obtained. Some further remarks are in order:

- Given $B \in \mathcal{F}$ such that $P(B) > 0$, the conditional probability $P(\cdot|B)$ is a probability measure on $(\Omega, \mathcal{F})$
- If the events $A, B \in \mathcal{F}$ are independent and such that $P(A) > 0$ and $P(B) > 0$, we have $P(A|B) = P(A)$ and $P(B|A) = P(B)$.
- Whenever well defined, $P(A|B) \neq P(B|A)$ in general. Indeed, Bayes theorem gives:

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)} \ .$$

We conclude this section with a trivial example. Consider again the random experiment consisting of tossing a coin twice. Assume the coin is balanced. It is clear that the probability of obtaining two heads is 1/4. Now, assume that we had already tossed the coin one time and that the outcome was head. What can we say about the probability of obtaining two heads? This is a very simple but typical example of conditional probability. With the extra available information, the probability of obtaining two heads is now 1/2. As an exercise, think of some (non-independent) pairs of events, and compute the corresponding conditional probabilities, always trying to obtain an interpretation for the results obtained.

2.2 Random variables

A random variable is a map a from the sample space Ω into the set of real numbers $\mathbb{R}$ (or $\mathbb{R}^K$, for some $K \in \mathbb{N}$, in the case of random vectors) satisfying a measurability condition to be detailed below. Its introduction allows one to move the study of a given random experiment from the sample space Ω to the set of real numbers, more suitable for the mathematical treatment of the problem at hand.

Before giving a precise definition of a random variable, we need to introduce the concept of measurable function.

Definition 2.2.1 (Measurable function)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space. A function $f : \Omega \to \mathbb{R}$ is* $\mathcal{F}$-measurable *(or simply measurable) if for each $a, b \in \mathbb{R}$ we have*

$$\{\omega \in \Omega : a \leq f(\omega) < b\} \in \mathcal{F} \ .$$

It should be remarked that the definition of measurable function given above, can be restated in a more general way. Start by noting that the definition above can be rewritten as follows: a function $f : \Omega \to \mathbb{R}$ is $\mathcal{F}$-measurable if $f^{-1}([a, b)) \in \mathcal{F}$, for each $a, b \in \mathbb{R}$. This should be interpreted in the following way: the relevant requirement in the definition above is that the preimage of every set in the Borel σ-algebra of $\mathbb{R}$ is in $\mathcal{F}$. More generally, we say that a function $f : \Omega \to \Omega'$ between two measurable spaces $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ is measurable if and only if for every set $B \in \mathcal{F}'$, its preimage $f^{-1}(B) \in \mathcal{F}$.

It is possible to check that linear combinations and products of measurable functions are again measurable functions. Moreover, if the sample space Ω is discrete and equipped with the σ-algebra consisting of all the subsets of Ω, then any real-valued function on Ω is measurable.

We are now ready to define what is meant by random variable.

Definition 2.2.2 (Random variable)**.** *A measurable function $f : \Omega \to \mathbb{R}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ is called a* random variable.

Similarly, a measurable function $f : \Omega \to \mathbb{R}^K$ defined on a probability space $(\Omega, \mathcal{F}, P)$ is called a *random vector*. From an heuristic point of view, a random variable $X : \Omega \to \mathbb{R}$ identifies the sample space Ω with a subset of the real numbers $X(\Omega) \subseteq \mathbb{R}$. There are several advantages in this approach. First of all, the introduction of a random variable enables one to move the analysis of a given random experiment to the set of real numbers (or some other multidimensional euclidean space), for which standard analytical techniques are readily available. Secondly, one is free to choose the most suitable random variable to analyse a particular random experiment. This is illustrated in the example below.

Example Recall the random experiments described above.

(i) Single coin toss. One possible random variable to assign to this random experiment would be

$$X(\omega) = \begin{cases} 1\ , & \text{if } \omega = H \\ 0\ , & \text{if } \omega = T \end{cases} .$$

An easy exercise: argue that X is a random variable.

(ii) Several coin tosses. One can use the random variable of the previous item to examine this random experiment. The sample space for this modified random experiment is $\Omega = \{H, T\}^n$, i.e. the space of sequences of two symbols (H and T) of length n:

$$\{H, T\}^n = \{(\omega_1, \omega_2, \ldots, \omega_n) : \omega_i \in \{H, T\}, i = 1, 2, \ldots, n\} .$$

Let X_i be the random variable on Ω defined by $X_i(\omega) = 1$ if $\omega_i = H$ (the i-th coin toss outcome is heads), and $X_i(\omega) = 0$ if $\omega_i = T$ (the ith coin toss outcome is tails). We define another random variable $Y : \Omega \to \mathbb{R}$ by

$$Y(\omega) = \sum_{i=1}^{n} X_i(\omega) .$$

The random variable Y maps each sequence $\omega \in \{H, T\}^n$ to the number of heads observed in such sequence.

(iii) Rolling a dice with six faces. Recall that the sample space for a single dice rolling is

$$\Omega = \{\text{“face 1”}, \text{“face 2”}, \ldots, \text{“face 6”}\} .$$

One can easily think about assigning the following random variable to this random experiment:

$$X(\omega) = i \quad \text{if } \omega = \text{“face i”} .$$

Consider now the experiment of rolling the dice twice. The sample space is $\Omega^2 = \{(\omega_1, \omega_2) : \omega_1, \omega_2 \in \Omega\}$ and one can consider the random variable $Y : \Omega \to \mathbb{R}^2$ defined by

$$Y(\omega_1, \omega_2) = (i, j) \quad \text{if } \omega_1 = \text{“face i”} \text{ and } \omega_2 = \text{“face j”} .$$

However, if one is only interested in the sum of points obtained in the two dice rollings, we could also consider the random variable $Z : \Omega^2 \to \mathbb{R}$ given by

$$Z(\omega_1, \omega_2) = i + j \quad \text{if } \omega_1 = \text{“face i”} \text{ and } \omega_2 = \text{“face j”} .$$

(iv) Rolling a dice until the outcome “face 6” is realized. Exercise: think of two distinct non-trivial random variables or vectors that may be assigned to this random experiment.

(v) The height of a randomly chosen 10 year old Portuguese child. The most natural random variable to assign to this particular random experiment is the function whose image is the numerical value of the height of the child for some choice of measurement units.

The next definition relates probability with the notion of random variable.

Definition 2.2.3 (Distribution function)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let X be a random variable on $(\Omega, \mathcal{F}, P)$. The* distribution function of the random variable X *is the function $F_X : \mathbb{R} \to \mathbb{R}$ given by*

$$F_X(x) = P(\{\omega \in \Omega : X(\omega) \leq x\}) , \qquad x \in \mathbb{R} .$$

Note that there are two functions involved in the definition of the distribution function F_X: the random variable X and the probability measure P. Hence, the distribution function carries with it information from both functions. Moreover, it is their joint use that enables the definition of the distribution function as a real function with a real variable. This is one particular feature of the distribution function that makes it so amenable to mathematical treatment. We list below some other properties satisfied by a distribution function of a random variable.

Theorem 2.2.4. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. If F_X is the distribution function of a random variable X on $(\Omega, \mathcal{F}, P)$, then*

1) F_X is non-decreasing, that is $F_X(x) \leq F_X(y)$ if $x \leq y$.

2) $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$.

3) $F_X(x)$ is continuous from the right for every $x \in \mathbb{R}$, that is:

$$\lim_{y \to x^+} F_X(y) = F_X(x) .$$

Any function $F : \mathbb{R} \to \mathbb{R}$ which satisfies the three properties listed in the previous theorem is called a *distribution function*. Moreover, any distribution function defines a probability measure on the set of real numbers $\mathbb{R}$ endowed with the corresponding Borel σ-algebra.

The concept of distribution function is easily generalized to several dimensions.

Definition 2.2.5 (Distribution function of a random vector)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $X : \Omega \to \mathbb{R}^K$ be a random vector on $(\Omega, \mathcal{F}, P)$. The* distribution function of the random vector X *is the function $F_X : \mathbb{R}^K \to \mathbb{R}$ given by*

$$F_X(x_1, \ldots, x_K) = P(\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_K(\omega) \leq x_K\}) \ ,$$

where $X_1(\omega), \ldots, X_K(\omega)$ are the K components of $X(\omega)$.

Let $x = (x_1, \ldots, x_K)$ and $y = (y_1, \ldots, y_K)$ be two vectors in $\mathbb{R}^K$. We say that $x \leq y$ if and if $x_i \leq y_i$ for every $i \in \{1, \ldots, n\}$. Similarly to the one-dimensional case, a distribution function of a random vector has the following properties:

1) F_X is non-decreasing with respect to the order in $\mathbb{R}^K$ defined above, i.e.

$$F_X(x) \leq F_X(y) \text{ if } x \leq y \ .$$

2) $\lim_{x \to (-\infty, \ldots, -\infty)} F_X(x) = 0$ and $\lim_{x \to (+\infty, \ldots, +\infty)} F_X(x) = 1$.

3) $F_X(x)$ is continuous from above for every $x \in \mathbb{R}^K$, that is:

$$\lim_{y \to x^+} F_X(y) = F_X(x) \ .$$

As before, any function $F : \mathbb{R}^K \to \mathbb{R}$ which satisfies the three properties listed above is called a *distribution function*.

We will now restrict our attention to two important classes of random variables: discrete and continuous. It is important to remark that these two cases are not mutually exclusive, i.e. there are random variables which do not fit into any of these classes. A complete treatment of this subject is based on the study of measure theory and is outside of the scope of this course. We discuss the case of discrete random variables first.

Definition 2.2.6 (Discrete random variable and probability function)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \mathcal{F}, P)$. The random variable X is said to be* discrete *if its distribution function $F_X : \mathbb{R} \to \mathbb{R}$ is a step function with a countable number of discontinuities.*
Denote by D_X the set of discontinuity points of F_X and define a function $f_X : \mathbb{R} \to \mathbb{R}$ through

$$f_X(x) = \begin{cases} F_X(x) - \lim_{y \to x^-} F_X(y) \ , & \text{if } x \in D_X \\ 0 \ , & \text{otherwise} \end{cases} \ .$$

The function f_X defined above is called probability function *of X.*

The probability function f_X of a discrete random variable X on a probability space $(\Omega, \mathcal{F}, P)$ has the following properties:

1) $f_X(x) > 0$ for every $x \in D_X$.

2) $\sum_{x \in D_X} f_X(x) = 1$.

3) $F_X(x) = \sum_{\{y \in D_x:\, y \leq x\}} f_X(y)$.

Example Recall the examples discussed above.

(i) Single coin toss. Assume that the coin is balanced and consider the random variable $X : \Omega \to \mathbb{R}$ given by

$$X(\omega) = \begin{cases} 1\ , & \text{if } \omega = H \\ 0\ , & \text{if } \omega = T \end{cases} .$$

Its distribution function is given by

$$F_X(\omega) = \begin{cases} 0\ , & \text{if } x < 0 \\ 1/2\ , & \text{if } 0 \leq x < 1 \\ 1\ , & \text{if } x \geq 1 \end{cases} .$$

Thus, X is a discrete random variable with probability function given by

$$f_X(x) = \begin{cases} 1/2\ , & \text{if } x = 0 \text{ or } x = 1 \\ 0\ , & \text{otherwise} \end{cases} .$$

(ii) n coin tosses. Similarly to what was done before, let X_i be the random variable on $\Omega = \{H, T\}^n$ defined by $X_i(\omega) = 1$ if $\omega_i = H$ and $X_i(\omega) = 0$ if $\omega_i = T$, $i \in \{1, \ldots, n\}$. Consider again the random variable $Y : \Omega \to \mathbb{R}$ defined by

$$Y(\omega) = \sum_{i=1}^{n} X_i(\omega)\ .$$

It should be clear that the random variable Y takes values on the set $\{0, 1, \ldots, n\}$ and is a discrete random variable. Exercise: why is this true?

Indeed, the random variable Y has a well-know distribution, that will be discussed in some more detail below. Assume that in a single coin toss the probability of obtaining heads is p, and the probability of obtaining tails is $q = (1 - p)$. Then, the probability function of Y is known to be given by

$$f_X(x) = \begin{cases} \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}\ , & \text{if } x \in \{0, 1, \ldots, n\} \\ 0\ , & \text{otherwise} \end{cases} .$$

The corresponding distribution function is

$$F_X(x) = \sum_{k=0}^{[x]} f_X(k)\ ,$$

where $[x]$ denotes the integer part of x.

(iv) Rolling a dice until the outcome "face 6" is realized. We have seen before that the sample space for this random experiment is

$$\Omega = \{\text{“1”}, \text{“2”}, \text{“3”}, \ldots, \text{“}\infty\text{”}\} .$$

One can define the random variable $X : \Omega \to \mathbb{R}$ given by

$$X(\text{“i”}) = i .$$

Under the assumption that the dice is balanced, we have already computed the probabilities of the elementary events of Ω. Indeed, this determines the probability function

$$f_X(x) = \begin{cases} \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{x-1} , & \text{if } x \in \mathbb{N} \\ 0 , & \text{otherwise} \end{cases}$$

and the corresponding distribution function

$$F_X(x) = \sum_{k=0}^{[x]} f_X(k) .$$

This particular probability distribution is known as geometric distribution.

We will now move on to the topic of continuous random variables.

Definition 2.2.7 (Continuous random variable and probability density)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \mathcal{F}, P)$. The random variable X is said to be* continuous *if its distribution function is a continuous function and there exists a non-negative integrable function $f : \mathbb{R} \to \mathbb{R}$ such that*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, \mathrm{d}t .$$

The function f_X is called probability density *of X.*

The probability density f_X of a continuous random variable X on a probability space $(\Omega, \mathcal{F}, P)$ has the following properties:

1) $\int_{-\infty}^{+\infty} f_X(x)\, \mathrm{d}x = 1$.
2) $P(X(\omega) \leq a) = \int_{-\infty}^{a} f_X(x)\, \mathrm{d}x$.

Example Two examples of probability distributions of continuous random variables.

(i) The height of a randomly chosen 10 year old Portuguese child. There is not a unique probability distribution model for this random experience. One simple choice for the distribution of this random variable is the normal distribution with properly chosen mean and variance. However, even though this may be usually a good approximation, it has the inconvenience of assigning a small positive probability to negative values of height.

(ii) Waiting time for some event to happen. Consider the following random experiment: given some piece of electronic equipment, measure the time it works before any malfunction. It should be clear that the sample space may be taken as $\Omega = \mathbb{R}_0^+$, or even $\Omega = \mathbb{R}$ for simplicity of treatment. Take for σ-algebra the Borel σ-algebra of $\mathbb{R}$. Define a random variable $X : \Omega \to \mathbb{R}$ by

$$X(\{\text{it takes } x \text{ units of time before malfunction}\}) = x \ .$$

A standard choice for the distribution function of this random variable is the exponential distribution. This is a one-parameter family of distributions with probability density given by

$$f(x;\lambda) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} \ , & x \geq 0 \\ 0 \ , & x < 0 \end{cases} \ ,$$

where $\lambda > 0$. The corresponding distribution function is given by

$$F(x;\lambda) = \begin{cases} 1 - \mathrm{e}^{-\lambda x} \ , & x \geq 0 \\ 0 \ , & x < 0 \end{cases} \ .$$

The parameter λ is related with the average waiting time before a malfunction, equal to $1/\lambda$.

Before moving on, we remark that the notion of distribution function of a random variable, as well as those of probability function and density function, are easily extend to the multidimensional setup of random vectors.

Definition 2.2.8 (Discrete random vector and probability function)**.** *Let $X : \Omega \to \mathbb{R}^K$ be a random vector on a probability space $(\Omega, \mathcal{F}, P)$. The random vector $X = (X_1, \ldots, X_K)$ is said to be* discrete *if each one of its components X_i, $i \in \{1, \ldots, K\}$, is a discrete random variable. The function $f : \mathbb{R}^K \to \mathbb{R}$ given by*

$$f_X(x_1, \ldots, x_K) = P(\{\omega \in \Omega : X(\omega) = (x_1, \ldots, x_K)\})$$

is called the probability function *of X.*

Note that the set of points $D_X = \{(x_1, \ldots, x_K) \in \mathbb{R}^K : f_X(x_1, \ldots, x_K) > 0\}$ is at most countable.

For the continuous case, we have the following definition.

Definition 2.2.9 (Continuous random vector and probability density)**.** *Let $X : \Omega \to \mathbb{R}^K$ be a random vector on a probability space $(\Omega, \mathcal{F}, P)$. The random vector X is said to be* continuous *if its distribution function is a continuous function and there exists a non-negative integrable function $f : \mathbb{R}^K \to \mathbb{R}$ such that*

$$F_X(x_1, \ldots, x_K) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_K} f_X(t_1, \ldots, t_K) \, \mathrm{d}t_K \ldots \mathrm{d}t_1 \ .$$

The function f_X is called probability density *of X.*

We will now introduce the concept of mathematical expectation, followed by a brief discussion of some other parameters useful for the description of probability distributions.

Definition 2.2.10 (Mathematical Expectation)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \mathcal{F}, P)$.*

- *If X is a discrete random variable, its* mathematical expectation *$E[X]$ is*

$$E[X] = \sum_{x \in D_X} x f_X(x) \ .$$

- *If X is a continuous random variable, its* mathematical expectation *$E[X]$ is*

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x) \, \mathrm{d}x \ .$$

Note that the mathematical expectation of a given random variable may not be finite, i.e. the sum or integral defining it may be divergent. If finite, the mathematical expectation provides a measure of localization for the distribution of a random variable X. Other terminology includes: expected value, mean, and mean value.

Proposition 2.2.11 (Properties of the Mathematical Expectation)**.** *Let X and Y be random variables with finite mathematical expectation on a probability space $(\Omega, \mathcal{F}, P)$ and let a, b, c be real numbers. The following properties hold:*

1) If $X(\omega) = c$ for every $\omega \in \Omega$, then $E[X] = c$.

2) The mathematical expectation of the random variable $aX + bY$ is finite and

$$E[aX + bY] = aE[X] + bE[Y] \ .$$

3) If $X \geq 0$ then $E[X] \geq 0$.

4) If $a \leq X \leq b$ then $a \leq E[X] \leq b$.

5) Chebyshev Inequality: *If $X \geq 0$, then for each $a > 0$ we have*

$$P(X \geq a) \leq \frac{E[X]}{a} \ .$$

Let X be a random variable on a probability space $(\Omega, \mathcal{F}, P)$ and let $Y = g(X)$, for some measurable function $g : \mathbb{R} \to \mathbb{R}$. Then, Y is also a random variable on $(\Omega, \mathcal{F}, P)$. If X is a discrete random variable, then Y is also discrete. In the case where X is a continuous random variable, Y is not necessarily continuous. The mathematical expectation of $Y = g(X)$ is given by

$$E[g(X)] = \sum_{x \in D_X} g(x) f_X(x)$$

in the case where X is a discrete random variable and

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) \, \mathrm{d}x$$

in the case where both X and Y are continuous random variables.

The definition above can be extended to functions of random vectors. Let $X : \Omega \to \mathbb{R}^K$ be a random vector on a probability space $(\Omega, \mathcal{F}, P)$ and let $Y = g(X)$ for some measurable function $g : \mathbb{R}^K \to \mathbb{R}$. Then, if X is discrete

$$E[g(X)] = \sum_{(x_1, \ldots, x_K) \in D_X} g(x_1, \ldots, x_K) f_X(x_1, \ldots, x_K)$$

and if X and Y are both continuous

$$E[g(X)] = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} g(x_1, \dots, x_K) f_X(x_1, \dots, x_K) \, \mathrm{d}x_1 \dots \mathrm{d}x_K \ .$$

In the case of measurable functions $g : \mathbb{R}^K \to \mathbb{R}^N$ of a random vector $X : \Omega \to \mathbb{R}^K$, one defines

$$E[g(X)] = (E[g_1(X)], \dots, E[g_N(X)]) \ ,$$

where the functions $g_i : \mathbb{R}^K \to \mathbb{R}$, $i \in \{1, \dots, N\}$, are the N components of g.

Based on the definition of mathematical expectation, we will now introduce a measure of dispersion – the variance. The variance measures how much a probability distribution spreads around its expected value.

Definition 2.2.12 (Variance)**.** *Let X be a random variable (or random vector) on a probability space $(\Omega, \mathcal{F}, P)$. The* variance *of X, denoted by* $\mathrm{Var}(X)$*, is equal to*

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

As with the mathematical expectation of a random variable, the variance may also not be finite. However, if $\mathrm{Var}(X)$ is finite, then so is $E[X]$, but the reciprocal statement is not true.

Proposition 2.2.13 (Properties of the Variance)**.** *Let X and Y be random variables with finite mathematical expectation on a probability space $(\Omega, \mathcal{F}, P)$ and let a, b, c be real numbers. The following properties hold:*

1) $\mathrm{Var}[X]$ *is finite if and only if $E[X^2]$ is finite. In this case*

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

2) If $X(\omega) = c$ for every $\omega \in \Omega$, then $Var[X] = 0$.

3) If $\mathrm{Var}[X]$ *is finite, the variance of $aX + b$ is finite and*

$$\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X] \ .$$

4) If $a \leq X \leq b$ then $\mathrm{Var}[X] \leq (b - a)^2/4$.

5) Chebyshev Inequality: *If* $\mathrm{Var}[X]$ *is finite, then for each $a > 0$ we have*

$$P\left(|X - E[X]| \geq a\right) \leq \frac{\mathrm{Var}[X]}{a^2} \ .$$

There are other interesting parameters describing the shape of a distribution, such as the skewness and kurtosis parameters. More information on this subject can be easily found in the references.

We will now introduce two quantities used to measure the linear dependence between two random variables: the covariance and the correlation coefficient, the latter being a dimensionless normalization of the covariance. If larger values of one random variable correspond to larger values of the second one, then the covariance is positive. On the other hand, if larger values of one random variable correspond to smaller values of the second one, then the covariance is negative.

Definition 2.2.14 (Covariance)**.** *Let X and Y be random variables on a probability space $(\Omega, \mathcal{F}, P)$. The* covariance *of X and Y, denoted by $\mathrm{Cov}[X, Y]$, is equal to*

$$\mathrm{Cov}[X, Y] = E\left[(X - E[X])\,(Y - E[Y])\right] \ .$$

Some properties of the covariance are listed below.

Proposition 2.2.15 (Properties of the Covariance)**.** *Let X and Y be random variables with finite variance on a probability space $(\Omega, \mathcal{F}, P)$ and let a, b, c, d be real numbers. The following properties hold:*

1) $\mathrm{Cov}[aX + b, cY + d] = ac\,\mathrm{Cov}[X, Y]$.

2) $(\mathrm{Cov}[X, Y])^2 \leq \mathrm{Var}[X]\mathrm{Var}[Y]$.

3) $\mathrm{Cov}[X, Y] = E[XY] - E[X]E[Y]$.

Definition 2.2.16 (Correlation coefficient)**.** *Let X and Y be random variables (or random vectors) with non-zero variance on a probability space $(\Omega, \mathcal{F}, P)$. The* correlation coefficient *of X and Y, denoted by $\rho[X, Y]$, is equal to*

$$\rho[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} \ .$$

Some properties of the correlation coefficient are listed below.

Proposition 2.2.17 (Properties of the correlation coefficient)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let X and Y be random variables on $(\Omega, \mathcal{F}, P)$ with non-zero variance. Let a, b, c, d be real numbers. The following properties hold:*

1) $-1 \leq \rho[X, Y] \leq 1$.

2) $\rho[aX + b, cY + d] = \rho[X, Y]$.

3) $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + 2\,\mathrm{Cov}[X, Y] + \mathrm{Var}[Y]$.

To end this section, we introduce and discuss the concept of independence of random variables

Definition 2.2.18 (Independent random variables)**.** *Let X and Y be random variables on a probability space $(\Omega, \mathcal{F}, P)$. We say that X and Y are* independent random variables *if for every $a, b \in \mathbb{R}$ the events $\{\omega \in \Omega : X(\omega) \leq a\}$ and $\{\omega \in \Omega : X(\omega) \leq b\}$ are independent.*

The definition above can be given in terms of σ-algebras, as we pass to explain. Let $\mathcal{F}_X \subseteq \mathcal{F}$ be the smallest sub-σ-algebra of $\mathcal{F}$ that makes X a measurable function. Define $\mathcal{F}_Y \subseteq \mathcal{F}$ in a similar fashion. Then, X and Y are independent if for any $A \in \mathcal{F}_X$ and any $B \in \mathcal{F}_Y$, the events A and B are independent. The notion of independence may also be expressed in terms of distribution or probability functions.

Theorem 2.2.19. *Let X and Y be random variables on a probability space $(\Omega, \mathcal{F}, P)$. Denote by F_X and F_Y the distribution functions of X and Y, respectively, and by f_X and f_Y the corresponding probability functions (densities in the continuous case). The following statements are equivalent:*

(i) X and Y are independent.

(ii) the distribution function of the random vector (X, Y) is such that

$$F_{(X,Y)}(x, y) = F_X(x)F_Y(y)$$

for every $x, y \in \mathbb{R}$.

(iii) the probability function (density) of the random vector (X, Y) is such that

$$f_{(X,Y)}(x, y) = f_X(x)f_Y(y)$$

for every $x, y \in \mathbb{R}$.

The following properties also hold.

Theorem 2.2.20. *Let X and Y be independent random variables on a probability space $(\Omega, \mathcal{F}, P)$. Then:*

(i) $E[XY] = E[X]E[Y]$.

(ii) $\mathrm{Cov}[X, Y] = 0$.

(iii) $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

2.3 Discrete distributions

This section is devoted to a brief description of some of the most common probability distributions for discrete random variables.

2.3.1 Discrete uniform

The discrete uniform distribution in n points, $n > 1$, is a probability distribution under which n distinct outcomes have equal probability. This distribution may be used to describe random experiments such as tossing a balanced coin or rolling a balanced dice.

Definition 2.3.1 (Discrete uniform distribution)**.** *We say that the random variable X follows a* discrete uniform distribution *in the set of points $\{x_1, \ldots, x_n\}$, and denote it by $X \sim U(x_1, \ldots, x_n)$, if its probability function is*

$$f(x) = \begin{cases} 1/n\ , & \text{if } x \in \{x_1, \ldots, x_n\} \\ 0\ , & \text{otherwise} \end{cases} .$$

The expected value and variance of a random variable $X \sim U(x_1, \ldots, x_n)$ are

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i\ , \qquad \mathrm{Var}[X] = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - (E[X])^2\ .$$

2.3.2 Bernoulli

The Bernoulli distribution is used to describe random experiments with two possible complementary outcomes: "success" (with probability p) when one given event A is observed and "failure" (with probability $1-p$) when that same event is not observed. Such random experiments are called *Bernoulli trials*. One example of a Bernoulli trial: rolling a dice to obtain "face 6". The outcome "success" is observing "face

6", while the outcome "failure" corresponds to the observation of any of the other five faces.

The sample space for the class of random experiments described above is $\Omega = \{S, F\}$, where S denotes the outcome "success" and F denotes the outcome "failure". Define the random variable

$$X(\omega) = \begin{cases} 1\ , & \text{if } \omega = S \\ 0\ , & \text{if } \omega = F \end{cases} .$$

Definition 2.3.2 (Bernoulli distribution)**.** *We say that X follows a* Bernoulli distribution*, and denote it by $X \sim Bi(1, p)$, if its probability function is*

$$f(x) = \begin{cases} p\ , & \textit{if } x = 1 \\ 1 - p\ , & \textit{if } x = 0 \\ 0\ , & \textit{otherwise} \end{cases} .$$

Note that the Bernoulli distribution is a one-parameter family of distributions depending on p, the probability of the outcome "success". The expected value and variance of a random variable $X \sim Bi(1, p)$ are

$$E[X] = p\ , \qquad \text{Var}[X] = p(1 - p)\ .$$

2.3.3 Binomial

The Binomial distribution describes the probability distribution of a random variable associated with the following random experiment: how many outcomes "success" are observed in a repetition of $n \geq 1$ independent Bernoulli trials with success probability $p \in (0, 1)$. Examples of this kind of random experiment include the number of "heads" observed in several coin tosses, or the number of "face 6" observed in several dice rollings.

The sample space for the class of random experiments described above is $\Omega = \{\text{“0”}, \text{“1”}, \ldots, \text{“n”}\}$. Define the random variable

$$X(\text{“i”}) = i\ .$$

Definition 2.3.3 (Binomial distribution)**.** *We say that X follows a* Binomial distribution *with parameters n and p, and denote it by $X \sim Bi(n, p)$, if its probability function is*

$$f(x) = \begin{cases} \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1 - p)^{n-x}\ , & \textit{if } x \in \{0, 1, \ldots, n\} \\ 0\ , & \textit{otherwise} \end{cases} .$$

Note that the Bernoulli distribution is a two-parameter family of distributions depending on n, the number of repetitions of a given Bernoulli trial, and p, the probability of the outcome "success" in each Bernoulli trial. The Bernoulli distribution may be seen as a particular case of the Binomial distribution when the number of repetitions is $n = 1$. We list below some other properties for the Binomial family of random variables.

- The expected value and variance of $X \sim Bi(n,p)$ are

$$E[X] = np \ , \qquad \mathrm{Var}[X] = np(1-p) \ .$$

- If $X \sim Bi(n,p)$ and $Y \sim Bi(m,p)$ are independent random variables, then $X + Y \sim Bi(n+m,p)$.
- If $X_1, \ldots, X_n \sim Bi(1,p)$ is a sequence of independent Bernoulli random variables, then

$$\sum_{i=1}^{n} X_i \sim Bi(n,p) \ .$$

2.3.4 Negative Binomial

The Negative Binomial distribution describes the probability distribution of a random variable associated with the following random experiment: how many outcomes "success" are observed in a repetition of independent Bernoulli trials with success probability $p \in (0,1)$ before a specified number $n \geq 1$ of "failure" occurs. An example of this kind of random experiment: number of times that the faces "1", "2", "3", "4" and "5" are observed before "face 6" is observed twice (note that "face 6" corresponds to the outcome "failure" in this example).

The sample space for the class of random experiments described above may be identified with $\Omega = \mathbb{N}_0$. Define the random variable

$$X(\text{number } i \text{ of "successes" before } n \text{ "failures"}) = i \ .$$

Definition 2.3.4 (Negative Binomial distribution)**.** *We say that X follows a* Negative Binomial distribution *with parameters n and p, and denote it by $X \sim NB(n,p)$, if its probability function is*

$$f(x) = \begin{cases} \binom{x+n-1}{x} p^x (1-p)^n \ , & \text{if } x \in \mathbb{N}_0 \\ 0 \ , & \text{otherwise} \end{cases} \ .$$

The Negative Bernoulli distribution is a two-parameter family of distributions depending on n, the number of "failure" to be observed on a sequence of Bernoulli trials, and p, the probability of the outcome "success" in each Bernoulli trial. It should be remarked that the definition given above may be extended to positive real values of n. The geometric distribution of parameter p coincides with the Negative Binomial distribution $NB(1, 1-p)$. Some other properties of the Negative Binomial distribution:

- The expected value and variance of $X \sim NB(n,p)$ are

$$E[X] = \frac{np}{1-p} \ , \qquad \mathrm{Var}[X] = \frac{np}{(1-p)^2} \ .$$

- If $X \sim NB(n,p)$ and $Y \sim NB(m,p)$ are independent random variables, then $X + Y \sim NB(n+m,p)$.
- If $X_1, \ldots, X_n \sim NB(1,p)$ is a sequence of independent Negative Bernoulli random variables, then

$$\sum_{i=1}^{n} X_i \sim NB(n,p) \ .$$

2.3.5 Hypergeometric

The Hypergeometric distribution describes the probability distribution of a random variable associated with the following random experiment: how many outcomes "success" are observed in n draws from a finite population of size N without replacement. Note that the Binomial distribution corresponds to a similar random experiment with replacement after each trial. Examples of this kind of random experiment include the number of red cards observed in several draws from a deck of 52 cards.

Let $N > 1$ denote the population size, $m \in \{0, 1, \ldots, N\}$ denote the number of elements of the population identified with the outcome "success" and $n \in \{0, 1, \ldots, N\}$ denote the number of draws without replacement taken from the population. The sample space for this class of random experiments is

$$\Omega = \{\max\{0, n + m - N\}, \ldots, \min\{n, m\}\} \ .$$

Define the random variable

$$X(\text{"i successes observed"}) = i \ .$$

Definition 2.3.5 (Hypergeometric distribution)**.** *We say that X follows a* Hypergeometric distribution *with parameters N, n and m, and denote it by $X \sim H(N, n, m)$, if its probability function is*

$$f(x) = \begin{cases} \dfrac{\binom{m}{x}\binom{M-m}{n-x}}{\binom{M}{n}} \ , & \text{if } x \in \{\max\{0, n + m - N\}, \ldots, \min\{n, m\}\} \\ 0 \ , & \text{otherwise} \end{cases} \ .$$

The Hypergeometric distribution is a three-parameter family of distributions. The expected value and variance of $X \sim H(N, n, m)$ are

$$E[X] = \frac{nm}{N} \ , \qquad \text{Var}[X] = \frac{nm(N-m)(N-n)}{N^2(N-1)} \ .$$

2.3.6 Poisson

The Poisson distribution describes the probability distribution of a random variable associated with the following random experiment: how many occurrences of a given event are observed during a fixed interval of time or space under the assumption that these events occur with a known average rate and independently of the time or space elapsed since the last occurrence. Examples of this kind of random experiment: number of ships entering a port over an hour, and the number of telephone calls going through a central during a five minutes interval.

The sample space for this class of random experiments may be identified with $\Omega = \mathbb{N}_0$. Define the random variable

$$X(\text{number } i \text{ of occurrences}) = i \ .$$

Definition 2.3.6 (Poisson distribution)**.** *We say that X follows a* Poisson distribution *with parameter $\lambda > 0$, and denote it by $X \sim Po(\lambda)$, if its probability function is*

$$f(x) = \begin{cases} \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} , & \text{if } x \in \mathbb{N}_0 \\ 0 , & \text{otherwise} \end{cases} .$$

The Poisson distribution is a one-parameter family of distributions depending on λ – the average rate of occurrence per unit of time or space of the particular event under observation. Further properties of the Poisson distribution:

- The expected value and variance of $X \sim Po(\lambda)$ are

$$E[X] = \lambda , \qquad \mathrm{Var}[X] = \lambda .$$

- If $X \sim Po(\lambda_1)$ and $Y \sim Po(\lambda_2)$ are independent random variables, then $X + Y \sim Po(\lambda_1 + \lambda_2)$.
- Let $X_1, \ldots, X_n$ be n independent Bernoulli trials. Assume that the probability of success p_n for the n Bernoulli trials depends on n is such a way that $\lim_{n\to\infty} np_n = \lambda > 0$. Then

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{n} X_i = x\right) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} ,$$

  that is, the random variable $\sum_{i=1}^{n} X_i$ is asymptotically Poisson distributed.

2.4 Continuous distributions

This section is devoted to a brief description of some of the most common probability distribution for continuous random variables.

2.4.1 Uniform

The Uniform distribution on an interval $[a, b] \subset \mathbb{R}$ describes the probability distribution of a random variable with outcomes on a bounded interval of $\mathbb{R}$, assigning equal probability to subintervals of $[a, b]$ of the same size. This distribution is particularly useful to produce pseudo-random sequences of numbers in a computer.

Before proceeding to the definition of continuous uniform distribution, we define the *indicator function* of a set A as

$$\chi_A(x) = \begin{cases} 1 , & \text{if } x \in A \\ 0 , & \text{if } x \notin A \end{cases} .$$

Definition 2.4.1 (Uniform distribution)**.** *We say that a random variable X follows a* Uniform distribution *on the interval $[a, b]$, with $a, b \in \mathbb{R}$, and denote it by $X \sim U([a, b])$, if its probability function is*

$$f(x) = \frac{1}{b-a} \chi_{[a,b]}(x) .$$

The expected value and variance of a random variable $X \sim U([a, b])$ are

$$E[X] = \frac{a+b}{2} , \qquad \mathrm{Var}[X] = \frac{(b-a)^2}{12} .$$

2.4.2 Normal

The normal distribution with its famous bell-shaped probability density is a central element to probability theory and statistics. Its importance is closely related with the following asymptotic property for the distribution of the sum of random variables: under rather mild conditions the distribution of these sums converge to a normal distribution as the number of terms in the sum grows. This is the content of the central limit theorem, to be discussed below in more detail.

Definition 2.4.2 (Normal distribution)**.** *We say that a random variable X follows a* Normal distribution *with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, and denote it by $X \sim N(\mu, \sigma^2)$, if its probability function is*

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

The Normal distribution is a two-parameter family of distributions depending on its expected value μ and variance σ^2. One particular feature of the Normal distribution is that it is rather suitable to analytical treatment. This is mainly due to a number of properties, including:

- The expected value and variance of $X \sim N(\mu, \sigma^2)$ are
$$E[X] = \mu \ , \qquad \mathrm{Var}[X] = \sigma^2 \ .$$
- The probability density function of the normal distribution is symmetric with respect to its expected value.
- If $X \sim N(\mu, \sigma^2)$, then
$$\frac{X - \mu}{\sigma} \sim N(0, 1) \ .$$
The distribution $N(0, 1)$ is called the *standard normal distribution.*
- Let $X_1, \ldots, X_n$ be a sequence of normally distributed random variables such that for each $i \in \{1, \ldots, n\}$ we have
$$E[X_i] = \mu_i \ , \qquad \mathrm{Var}[X_i] = {\sigma_i}^2 \ , \qquad \mathrm{Cov}[X_i, X_j] = \sigma_{ij} \ .$$
Then, for every choice of constants $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, the random variable $\sum_{i=1}^n \alpha_i X_i$ is normally distributed, i.e. we have that
$$\sum_{i=1}^n \alpha_i X_i \sim N(\mu, \sigma^2) \ ,$$
where
$$\mu = \sum_{i=1}^n \alpha_i \mu_i \ , \qquad \sigma^2 = \sum_{i=1}^n {\alpha_i}^2 {\sigma_i}^2 + 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ j>i}}^n \alpha_i \alpha_j \sigma_{ij} \ .$$
- A particular consequence of the last statement is that if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ are independent random variables, then
$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2)$$
and
$$\frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right) \ .$$

2.4.3 Exponential

The Exponential distribution describes the probability distribution of the random variable associated with the following random experiment: measure the interval of time between two consecutive occurrences of a given event under the assumption that such event occurs with a known average rate and independently of the time elapsed since the last occurrence.

Definition 2.4.3 (Exponential distribution)**.** *We say that a random variable X follows an* Exponential distribution *with parameter $\lambda > 0$, and denote it by $X \sim Ex(\lambda)$, if its probability function is*

$$f(x) = \lambda \mathrm{e}^{-\lambda x} \chi_{\mathbb{R}_0^+}(x) \ .$$

The Exponential distribution is a one-parameter family of distributions depending on λ – the average rate of occurrence per unit of time or space of the particular event under observation. Clearly, it is closely related with the Poisson distribution. Further properties of the Exponential distribution:

- The expected value and variance of $X \sim Ex(\lambda)$ are

$$E[X] = \frac{1}{\lambda} \ , \qquad \mathrm{Var}[X] = \frac{1}{\lambda^2} \ .$$

- If $X \sim Ex(\lambda)$, then

$$P\left(X > x + y \,|X > y\right) = P(X > x)$$

  for every $x, y > 0$.

2.4.4 Gamma

The Gamma distribution is an important two-parameter family of continuous probability distributions, which includes the exponential family, the chi-square distribution, and the Erlang distribution.

Before proceeding to the definition of the Gamma distribution, we define the *gamma function* $\Gamma : \mathbb{R}^+ \to \mathbb{R}$ by

$$\Gamma(x) = \int_0^{+\infty} \mathrm{e}^{-x} x^{\alpha - 1} \, \mathrm{d}x \ .$$

The gamma function generalizes factorial to positive real numbers:

- $\Gamma(1) = 1$.
- If $n \in \mathbb{N}$, then $\Gamma(n) = (n-1)!$.
- $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, $\alpha > 1$.

Definition 2.4.4 (Gamma distribution)**.** *We say that a random variable X follows a* Gamma distribution *with parameters $\lambda, \alpha > 0$, and denote it by $X \sim Gamma(\alpha, \lambda)$, if its probability function is*

$$f(x) = \frac{\lambda^\alpha \mathrm{e}^{-\lambda x} x^{\alpha - 1}}{\Gamma(\alpha)} \chi_{\mathbb{R}_0^+}(x) \ .$$

Some properties of the Gamma distribution:

- The expected value and variance of $X \sim Gamma(\alpha, \lambda)$ are

$$E[X] = \frac{\alpha}{\lambda} \ , \qquad \mathrm{Var}[X] = \frac{\alpha}{\lambda^2} \ .$$

- If $X \sim Gamma(\alpha_1, \lambda)$ and $Y \sim Gamma(\alpha_2, \lambda)$ are independent random variables, then $X + Y \sim Gamma(\alpha_1 + \alpha_2, \lambda)$.
- The case $\alpha = 1$ corresponds to the exponential distribution.
- The case $\alpha = n \in \mathbb{N}$ corresponds to the Erlang distribution (which generalizes the exponential distribution). It describes the following random experiment: measure the interval of time elapsed before the observation of n consecutive occurrences of a given event under the assumption that such event occurs with a known average rate and independently of the time elapsed since the last occurrence.

2.4.5 Chi-square

The Chi-square distribution is yet another member of the Gamma family of probability distributions. It plays a prominent role in statistical inference.

Definition 2.4.5 (Chi-square distribution)**.** *We say that a random variable X follows a* Chi-square distribution *with n degrees of freedom, and denote it by $X \sim \chi^2(n)$, if its probability function is*

$$f(x) = \frac{\mathrm{e}^{-\frac{x}{2}} x^{\frac{n}{2}-1}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} \chi_{\mathbb{R}_0^+}(x) \ .$$

Some properties of the Chi-square distribution:

- A Chi-square distribution with n degrees of freedom is a Gamma distribution of the form $Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$
- The expected value and variance of $X \sim \chi^2(n)$ are

$$E[X] = n \ , \qquad \mathrm{Var}[X] = 2n \ .$$

- If $X \sim \chi^2(n_1)$ and $Y \sim \chi^2(n_2)$ are independent random variables, then $X + Y \sim \chi^2(n_1 + n_2)$.
- If $X \sim N(0, 1)$, then $X^2 \sim \chi^2(1)$.
- If $X_1, \ldots, X_n \sim N(0, 1)$ are independent random variables, then

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n) \ .$$

- If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ are independent random variables, then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n) \ .$$

2.4.6 Student's t

The Student's t distribution is related with the normal and chi-square distributions. Its important role in statistical inference is due to the following relation with the Normal and Chi-square distributions. Let U, V be independent random variables with distributions $U \sim N(0,1)$ and $V \sim \chi^2(n)$. Then, the random variable

$$\frac{U}{\sqrt{V/n}}$$

follows a Student's t distribution with n degrees of freedom.

Definition 2.4.6 (Student's t distribution)**.** *We say that a random variable X follows a* Student's t distribution *with n degrees of freedom, and denote it by $X \sim t(n)$, if its probability function is*

$$f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} .$$

Some properties of the Student's t distribution:

- The expected value and variance of $X \sim t(n)$ are

$$E[X] = 0 , \qquad \mathrm{Var}[X] = \frac{n}{n-2} .$$

- The probability density is symmetric with respect to its expected value.
- The particular case $n = 1$ is known as Cauchy's distribution.
- As $n \to \infty$, the Student's t probability density converges to the standard normal distribution probability density.

2.4.7 Snedecor's F distribution

The Snedecor's F distribution is widely used in statistical inference. Let U, V be independent random variables with distributions $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$. Then, the random variable

$$\frac{U/m}{V/n}$$

follows a Snedecor's F distribution distribution with m and n degrees of freedom.

Definition 2.4.7 (Snedecor's F distribution)**.** *We say that a random variable X follows a* Snedecor's F distribution distribution *with m and n degrees of freedom, and denote it by $X \sim F(m,n)$, if its probability function is*

$$f(x) = \frac{1}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}} \chi_{\mathbb{R}_0^+}(x) ,$$

where B denotes the Beta function:

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1} , \qquad \mathrm{Re}(x), \mathrm{Re}(y) > 0 .$$

Some properties of the Snedecor's F distribution:

- The expected value and variance of $X \sim F(m,n)$ are

$$E[X] = \frac{n}{n-2} \ , \qquad \text{for } n > 2$$

and

$$\mathrm{Var}[X] = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \ , \qquad \text{for } n > 4 \ .$$

- If $X \sim F(m,n)$, then $1/X \sim F(n,m)$.

2.5 The Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of random variables with finite mathematical expectation on a probability space $(\Omega, \mathcal{F}, P)$. Denote the mathematical expectation of each X_i, $i \in \mathbb{N}$, by $\mu_i = E[X_i]$. Moreover, denote by $\overline{X}_n$ and M_n the following averages:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \qquad M_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i \ .$$

Theorem 2.5.1 (Law of large numbers)**.** *A sequence $\{X_i\}_{i\in\mathbb{N}}$ of independent identically distributed random variables with finite mathematical expectation satisfies the Law of Large Numbers, i.e.*

$$P\left(\left|\overline{X}_n - \mu\right| > \epsilon\right) \to 0 \ , \quad \textit{as } n \to \infty \textit{ for any } \epsilon > 0.$$

Indeed, it is possible to drop the assumption that the random variables $\{X_i\}_{i\in\mathbb{N}}$ are identically distributed. If $\{X_i\}_{i\in\mathbb{N}}$ is a sequence of independent random variables with finite variance, i.e. there exists $\sigma > 0$ such that for all $i \in \mathbb{N}$ the variance of X_i is such that $\mathrm{Var}[X_i] < \sigma^2$, then the Law of Large Numbers still holds:

$$P\left(\left|\overline{X}_n - M_n\right| > \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2 n} \to 0 \ , \quad \text{as } n \to \infty \text{ for any } \epsilon > 0.$$

In the particular case of a sequence of homogeneous independent trials, the Law of Large Numbers states that typical realizations are such that the frequency with which an event occurs is close to the probability of this event. More precisely, let $\Omega = \{x_1, \ldots, x_k\}$ be a finite set with a probability measure P and let

$$p_j = P(x_j) \ , \qquad 1 \leq j \leq k \ .$$

Denote by ν_j^n the number of occurrences of the event x_j in a sequence of n independent trials. Then, for each $j \in \{1, \ldots, k\}$ we have that

$$P\left(\left|\frac{\nu_j}{n} - p_j\right| < \epsilon\right) \to 1 \ , \quad \text{as } n \to \infty.$$

Finally, we remark that the Law of large numbers concerns the convergence in probability of the sequence of random variables $\{\overline{X}_n\}_{n\in\mathbb{N}}$. Under stricter sets of hypotheses, stronger results hold such as the Strong Law of large numbers which ensures that the convergence is pointwise for almost every $\omega \in \Omega$.

2.6 The Central Limit Theorem

The central limit theorem is a key result in probability theory. It partially explains the huge relevance that the normal distribution has in mathematics. Very roughly, it states that the average

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

of a sequence of independent and identically distributed random variables $\{X_i\}_{i\in\mathbb{N}}$ converges in distribution to a normal distribution, no matter what the initial distribution of the initial random variables was.

We state below one version of Central limit theorem. The assumptions of the theorem can be relaxed to obtain stronger versions.

Theorem 2.6.1 (Central Limit theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite mean $\mu = E[X_i]$ and variance and $0 < \sigma^2 = Var[X_i]$, $i \in \mathbb{N}$. Then, the distribution functions of the random variables*

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

converge to the distribution function of the standard normal distribution $N(0,1)$ as $n \to \infty$.

For a sequence of independent Bernoulli trials with success probability p, we obtain the following result.

Theorem 2.6.2 (de Moivre-Laplace theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent Bernoulli distributed random variables with mean $p = E[X_i]$. Then, the distribution functions of the random variables*

$$Z_n = \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}}$$

converge to the distribution function of the standard normal distribution $N(0,1)$ as $n \to \infty$.

A consequence of the de Moivre-Laplace theorem is that for large enough n, the Normal distribution provides a good approximation for the Binomial distribution.

3 Stochastic Processes

A stochastic process is a collection of random variables $\{X_t\}_{t\in\mathbb{T}}$ on a probability space $(\Omega, \mathcal{F}, P)$ whose index $t \in \mathbb{T}$ is usually referred to as "time" when $\mathbb{T}$ is a subset of $\mathbb{R}$. In such case, stochastic processes describe the evolution with "time" of random phenomena. In this section we will provide the key definitions in this subject, and give a brief overview of the main properties of the following families of stochastic processes: Poisson process, Markov process and Brownian motion.

3.1 Basic properties

We start by providing the precise definition of a stochastic process.

Definition 3.1.1 (Stochastic process)**.** *A* stochastic process *is a collection of random variables $X = \{X_t : t \in \mathbb{T}\}$ on probability space $(\Omega, \mathcal{F}, P)$ taking values on a measurable space $(\Pi, \mathcal{G})$, indexed by a parameter t on a totally ordered set $\mathbb{T}$.*
The space $(\Omega, \mathcal{F}, P)$ is called sample space*, while $(\Pi, \mathcal{G})$ is called the* state space*. For a fixed sample point $\omega \in \Omega$, the function $t \to X_t(\omega), t \in \mathbb{T}$, is the* sample path *of the process X associated with ω.*

For simplicity of exposition, we will assume that the state space is $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-algebra of $\mathbb{R}^d$. If the set $\mathbb{T}$ is a subset of $\mathbb{R}$, usually $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}^+$ or $\mathbb{R}$, we think of the index $t \in \mathbb{T}$ as time.

Note that implicit in the definition of a stochastic process $X = \{X_t : t \in \mathbb{T}\}$ as a collection of $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$-valued random variables on $(\Omega, \mathcal{F}, P)$, is the assumption that each random variable X_t is $\mathcal{F}$-measurable (as discussed in the section concerning random variables). However, since X is a function of the pair of variables $(t, \omega) \in \mathbb{T} \times \Omega$, it is convenient to have joint measurability properties.

Definition 3.1.2 (Measurable stochastic process)**.** *Let $X = \{X_t : t \in \mathbb{T}\}$ be a stochastic process on the probability space $(\Omega, \mathcal{F}, P)$. The stochastic process X is called* measurable *if for every $A \in \mathcal{B}(\mathbb{R}^d)$ the set $\{(t, \omega) : X_t(\omega) \in A\}$ belongs to $\mathcal{B}(\mathbb{T}) \otimes \mathcal{F}$, i.e.*

$$X_t(\omega) : (\mathbb{T} \times \Omega, \mathcal{B}(\mathbb{T}) \otimes \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$$

is a measurable function.

The temporal feature of a stochastic process suggests a flow of time, in which, at every moment $t \in \mathbb{T}$, we can talk about past, present and future. In particular, we can ask how much an observer of the process knows about it at the present time, as compared to how much he knew at some point in the past or will know at some point in the future. The notion of σ-algebra is used in the study of stochastic processes to keep track of information as time evolves through the introduction of a filtration – a nested sequence of σ-algebras.

From now on, we assume that our sample space $(\Omega, \mathcal{F})$ is equipped with a filtration.

Definition 3.1.3 (Filtration)**.** *A* filtration *on a measurable space $(\Omega, \mathcal{F})$ is a non-decreasing family $\{\mathcal{F}_t : t \in \mathbb{T}\}$ of sub-σ-algebras of $\mathcal{F}$, i.e. $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$ for $s, t \in \mathbb{T}$ such that $s < t$.*

If $\mathbb{T}$ is an infinite set, we define $\mathcal{F}_\infty = \sigma\left(\bigcup_{t\in\mathbb{T}} \mathcal{F}_t\right)$ to be the smallest σ-algebra containing $\bigcup_{t\in\mathbb{T}} \mathcal{F}_t$.

The concept of measurability for a stochastic process introduced before is still rather weak. The introduction of a filtration $\{\mathcal{F}_t\}$ enables us to use more interesting and useful concepts.

Definition 3.1.4 (Adapted stochastic process)**.** *Let $X = \{X_t : t \in \mathbb{T}\}$ be a stochastic process on the probability space $(\Omega, \mathcal{F}, P)$. The stochastic process X is* adapted *to the filtration $\{\mathcal{F}_t\}$ if, for every $t \in \mathbb{T}$, X_t is an $\mathcal{F}_t$-measurable random variable.*

For a given stochastic process $X = \{X_t : t \in \mathbb{T}\}$ on a probability space $(\Omega, \mathcal{F}, P)$, the simplest choice for a filtration is the filtration generated by the process itself:

$$\mathcal{F}_t^X = \sigma(X_s : s \in \mathbb{T}\ , s \le t)\ ,$$

the smallest σ-algebra with respect to which X_s is measurable for every $s \in \mathbb{T}$ such that $s \le t$. It should be noted that every stochastic process X is adapted to the filtration $\{\mathcal{F}_t^X\}$.

A filtration can be seen as representing the flow of information. The σ-algebra $\mathcal{F}_t^X$ contains only the events that "can happen up to time t", i.e. when $A \in \mathcal{F}_t^X$, an observer of X during the time period $[0, t]$ knows whether or not the event A has occurred up to time t, but not after time t. Thus, an adapted process is one that "does not look into the future".

We will introduce a general notion of independency that will be useful in the sequel.

Definition 3.1.5 (Independent σ-algebras)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_n$ be sub-σ-algebras of $\mathcal{F}$. A finite set of sub-σ-algebras $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_n$ is* independent *if for any set of events $A_i \in \mathcal{F}_i$, $i = 1, ..., n$, we have that*

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)P(A_2)...P(A_n)\ .$$

An arbitrary set $\mathcal{S}$ of σ-algebras is mutually independent *if any finite subset of $\mathcal{S}$ is independent.*

The above definition is a generalization of the notions of independence for events and random variables:

- Events $B_1, ..., B_n \in \mathcal{F}$ are mutually independent if the sub-σ-algebras $\sigma(B_i) := \{\emptyset, B_i, \Omega - B_i, \Omega\}$ are mutually independent.
- Random variables $X_1, ..., X_n$ defined on $(\Omega, \mathcal{F}, P)$ are mutually independent if the sub-σ-algebras $\sigma(X_i) = \{X_i^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^d)\}$ are mutually independent.
- In general, mutual independence among events B_i, random variables X_j and σ-algebras $\mathcal{F}_k$ means the mutual independence among $\sigma(B_i)$, $\sigma(X_j)$ and $\mathcal{F}_k$.

Similarly to what was done before for events, one can define probability and mathematical expectation conditioned on a σ-algebra of events.

Definition 3.1.6 (Conditional expectation with respect to a σ-algebra)**.** *Let X be a random variable on the probability space $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{R}^d$, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The* conditional expectation of X given $\mathcal{G}$ *is denoted by $E[X|\mathcal{G}]$ and defined as the random variable on the probability space $(\Omega, \mathcal{G}, P)$ satisfying*

$$E\left[E[X|\mathcal{G}]\chi_A\right] = E[X\chi_A]\ , \qquad \textit{for all } A \in \mathcal{G}.$$

Let X, Y be random variables with finite mathematical expectation on the probability space $(\Omega, \mathcal{F}, P)$ and let $\alpha, \beta \in \mathbb{R}$. The conditional expectation with respect to the sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$ has the following properties:

- $E[\alpha X + \beta Y|\mathcal{G}] = \alpha E[X|\mathcal{G}] + \beta E[X|\mathcal{G}]$.
- $E[E[X|\mathcal{G}]] = E[X]$.
- $E[X|\mathcal{G}] = X$ if X is $\mathcal{G}$-measurable.
- $E[X|\mathcal{G}] = E[X]$ if X is independent of $\mathcal{G}$.
- $E[YX|\mathcal{G}] = YE[X|\mathcal{G}]$ if Y is $\mathcal{G}$-measurable.
- If $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{F}$ such that $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$, then

$$E[E[X|\mathcal{H}]|\mathcal{G}] = E[X|\mathcal{G}] \ .$$

Building up on the concept of conditional expectation, we define conditional probability with respect to a σ-algebra.

Definition 3.1.7 (Conditional probability with respect to a σ-algebra)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $A \in \mathcal{F}$ an event and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. The* conditional probability of A given $\mathcal{G}$ *is the conditional expectation of the indicator function of A, χ_A, given the sub-σ-algebra $\mathcal{G}$, that is*

$$P[A|\mathcal{G}] = E[\chi_A|\mathcal{G}] \ .$$

Similarly, we define the conditional probability of A given a random variable X *on $(\Omega, \mathcal{F})$ as*

$$P[A|X] = E[\chi_A|\sigma(X)] \ ,$$

where $\sigma(X)$ is the σ-algebra generated by the random variable X.

We now define Martingale: a stochastic process for which the expected future state is its current state.

Definition 3.1.8 (Martingale)**.** *Let $X = \{X_t : t \in \mathbb{T}\}$ be a stochastic process defined on a probability space $(\Omega, \mathcal{F}, P)$, adapted to a filtration $\{\mathcal{F}_t\}$. Furthermore, assume that $E|X_t| < \infty$ for all $t \in \mathbb{T}$. The process X is a* martingale *if, for every $s, t \in \mathbb{T}$ such $t \geq s$, we have*

$$E[X_t|\mathcal{F}_s] = X_s \ .$$

In the next sections we will discuss some special examples of stochastic processes.

3.2 Poisson Process

Fix a probability space $(\Omega, \mathcal{F}, P)$ and an event $A \in \mathcal{F}$ and count the number of times A occurs during an interval of time $[0, t]$, $t > 0$. Think of examples such as the arrival of ships to a port, the arrival of customers to a store, or the arrival of phone calls to a call center, during a certain period of time.

Denote by X_1 the time between $t = 0$ and the first occurrence of the event A, X_2 the time between the first and the second occurrences of the event A, and so on. Assume that the random variables $X_1, X_2, \ldots$ are independent and identically distributed. Additionally, assume that if the event A has not occurred until time

$t > 0$, then the distribution of the time remaining until the next occurrence of A is the same as the distribution of each of X_i, that is

$$P(X_i - t \in B | X_i > t) = P(X_i \in B)$$

for any Borel set $B \in \mathcal{B}(\mathbb{R})$. It is possible to show that if an unbounded random variable satisfies the property above, then it has exponential distribution.

Define a stochastic process $N : [0, \infty) \times \Omega \to \mathbb{R}$ by

$$N_t(\omega) = \sup \left\{ n \in \mathbb{N} : \sum_{i \leq n} X_i(\omega) \leq t \right\} .$$

Note that $N_t(\omega)$ is equal to the number of occurrences of the event A until time $t > 0$. Thus, we obtain that $N_t(\omega) \in \{0, 1, 2, \ldots\}$ for every $t > 0$ and every $\omega \in \Omega$.

Assume that the event A occurs with a known average rate $\lambda > 0$ per unit time. We have that:

- for each fixed $t > 0$, N_t is a random variable with Poisson distribution $Po(\lambda t)$;
- the waiting time for N_t to increase by one unit is a random variable with exponential distribution $Ex(\lambda)$;
- the waiting time for N_t to increase by n units is a random variable with Gamma distribution $Gamma(n, \lambda)$.

The formal definition of a Poisson process is the following.

Definition 3.2.1 (Poisson process)**.** *A stochastic process $N = \{N_t : t \in [0, \infty)\}$ on a probability space $(\Omega, \mathcal{F}, P)$ is called a Poisson process with parameter $\lambda > 0$ if the following properties hold:*

1) $N_0 = 0$ almost surely, i.e. $P(\omega \in \Omega : N_0(\omega) = 0) = 1$.

2) N_t is a process with independent increments, i.e. for $0 \leq t_1 \leq \ldots \leq t_k$ the random variables $N_{t_1}, N_{t_2} - N_{t_1}, \ldots, N_{t_k} - N_{t_{k-1}}$ are independent.

3) For any $0 \leq s < t < \infty$, the random variable $N_t - N_s$ has a Poisson distribution with parameter $\lambda(t - s)$.

We now list some further properties of a Poisson process:

- the sample path $t \to N_t$ is piecewise constant and continuous from the right;
- in points of discontinuity, the sample path jumps have unit size, i.e.

$$N_t - \lim_{s \to t^-} N_s \in \{0, 1\}$$

  for all $t > 0$.

- the stochastic process $N_t - \lambda t$ is a martingale.

An interesting situation, not yet covered by the notion of Poisson process as defined above, is the case of a stochastic process which shares most properties of the Poisson process except for the size of the increments, which are now allowed to be non-unitary. This is the case of a compound Poisson process. This may be used to model the amount of goods sold in a shop over a given period of time, or the number of containers arriving at a port on a given day.

Let N_t be a Poisson process with parameter $\lambda > 0$ on a probability space $(\Omega, \mathcal{F}, P)$ and let $\{J_i\}_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables on the same probability space with distribution function F. Assume that N_t and the random variables $\{J_i\}$ are independent. The stochastic process Y on $(\Omega, \mathcal{F}, P)$ given by

$$Y_t(\omega) = \sum_{i=1}^{N_t(\omega)} J_i(\omega)$$

is called a *compound Poisson process* with rate λ and jump distribution F. It satisfies the following properties:

- $Y_0 = 0$ almost surely.
- Y_t has independent increments.
- the number of increments of Y_t is Poisson distributed.
- the waiting time for the next increment is exponentially distributed.
- the size of the increments follows a probability distribution determined by the distribution function F.

3.3 Brownian motion

Brownian motion was first observed in 1827 by the biologist Robert Brown when looking through a microscope at pollen grains suspended in water. He observed that the pollen grains seem to move with a certain degree of randomness. This behaviour may be modeled by a stochastic process that is also know as Brownian motion.

The first person to obtain a mathematical model for Brownian motion seems to have been Louis Bachelier, a student of Henri Poincaré, in 1900. Bachelier interest in Brownian motion was connected with the time evolution of financial assets. In 1905, Albert Einstein studied Brownian motion from a physical perspective. The rigorous definition and the first mathematical proof of the existence of Brownian motion are due to the American mathematician Norbert Wiener in 1920.

Definition 3.3.1 (standard, one-dimensional Brownian motion)**.** *A stochastic process $B = \{B_t : 0 \le t < \infty\}$ on a probability space $(\Omega, \mathcal{F}, P)$ adapted to a filtration $\mathcal{F}_t$ is called a* standard, one-dimensional Brownian motion *if the following properties hold:*

1) the sample paths of B are continuous functions of t for almost all $\omega \in \Omega$;

2) $B_0 = 0$ almost surely;

3) for $0 \le s < t$, the increment $B_t - B_s$ is independent of $\mathcal{F}_s$;

4) for $0 \le s < t$, the increment $B_t - B_s$ is normally distributed with mean zero and variance $t - s$.

Analogously, we can define a Brownian motion $B = \{B_t : 0 \le t \le T\}$ on the interval $[0, T]$, for some $T > 0$.

If B is a Brownian motion and $0 = t_0 < t_1 < ... < t_n < \infty$, then the increments $B_{t_j} - B_{t_{j-1}}$, $j = 1, \ldots, n$, are independent and the distribution of $B_{t_j} - B_{t_{j-1}}$ depends on t_j and t_{j-1} only through the difference $t_j - t_{j-1}$: it is normal with mean zero and variance $t_j - t_{j-1}$. We say that B has stationary, independent increments.

Note that the filtration $\{\mathcal{F}_t\}$ is a key part in the definition of Brownian motion. However, if we are given $\{B_t : 0 \le t < \infty\}$ but no filtration, and if we know that B has stationary independent increments and that $B_t - B_0$ is normal with mean zero and variance t, we can take $\mathcal{F}_t^B = \sigma(B_t : 0 \le s \le t)$ for filtration. If $\{\mathcal{F}_t\}$ turns out to be "larger" than $\{\mathcal{F}_t^B\}$ (in the sense that $\mathcal{F}_t^B \subset \mathcal{F}_t$ for all $t \ge 0$) and if $B_t - B_s$ is independent of $\mathcal{F}_s$ whenever $0 \le s < t$, then $\{B_t : 0 \le t < \infty\}$ is still a Brownian motion with respect to $\mathcal{F}_t$.

Definition 3.3.2 (d-dimensional Brownian motion with initial distribution μ)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space equipped with a filtration $\{\mathcal{F}_t\}$, μ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, and d be a positive integer. A stochastic process $B = \{B_t : t \ge 0\}$ on $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{R}^d$ and adapted to $\{\mathcal{F}_t\}$ is called a* d-dimensional Brownian motion with initial distribution μ *if the following properties hold:*

1) the sample paths of B are continuous functions of t for almost all $\omega \in \Omega$;

2) $P[B_0 \in \Gamma] = \mu(\Gamma)$, for all $\Gamma \in \mathcal{B}(\mathbb{R}^d)$;

3) for $0 \le s < t$, the increment $B_t - B_s$ is independent of $\mathcal{F}_s$;

4) for $0 \le s < t$, the increment $B_t - B_s$ is normally distributed with mean zero and covariance matrix equal to $(t-s)I_d$, where I_d denotes the $d \times d$ identity matrix.

If μ assigns measure one to some singleton $\{x\}$, we say that B is a d-dimensional Brownian motion starting at x.

The following properties hold:

- Standard Brownian motion is a martingale.
- *Nowhere differentiability*: for almost every $\omega \in \Omega$, the Brownian sample path $B_t(\omega)$ is nowhere differentiable with respect to t.
- *Strong Law of Large Numbers*:

$$\lim_{t\to\infty} \frac{B_t}{t} = 0 \qquad \text{almost surely.}$$

- *Equivalence transformations.* If $B = \{B_t : 0 \le t < \infty\}$ is a standard Brownian motion, so are the processes obtained from the following equivalence transformations:
  - Scaling: $X = \{X_t : 0 \le t < \infty\}$ defined for $c > 0$ by

$$X_t = \frac{1}{\sqrt{c}} B_{ct} \ , \qquad 0 \le t < \infty \ .$$

  - Time-inversion: $Y = \{Y_t : 0 \le t < \infty\}$ defined by

$$Y_t = tB_{1/t} \ , \ 0 < t < \infty \ , \qquad Y_0 = 0 \ .$$

  - Time-reversal: $Z = \{Z_t : 0 \le t \le T\}$ defined for $T > 0$ by

$$Z_t = B_T - B_{T-t} \ , \ 0 \le t \le T \ .$$

  - Symmetry: $-B = \{-B_t : 0 \le t < \infty\}$.

3.4 Lévy process

Lévy processes form a large family of stochastic processes which includes, for instance, Brownian motion and the Poisson process. It is an adapted stochastic process with independent and stationary increments. The formal definition is given below.

Definition 3.4.1 (Lévy process)**.** *A stochastic process $X = \{X_t : 0 \leq t < \infty\}$ on a probability space $(\Omega, \mathcal{F}, P)$ adapted to a filtration $\mathcal{F}_t$ is called a* Lévy process *if the following properties hold:*

1) for $0 \leq s < t$, the increment $X_t - X_s$ is independent of $\mathcal{F}_s$;

2) X has stationary increments, that is, $X_t - X_s$, has the same distribution as X_{t-s}, $0 \leq s < t$;

3) X_t is continuous in probability, that is, for all $\epsilon > 0$

$$\lim_{s \to t} P\left(\omega \in \Omega : |X_s(\omega) - X_t(\omega)| \geq \epsilon\right) = 0 \ .$$

Examples of Lévy processes:

- A Poisson process with intensity $\lambda > 0$.
- A compound Poisson process with intensity $\lambda > 0$ and jumps distribution F.
- Let B be a standard Brownian motion, $\mu \in \mathbb{R}$ and $\sigma > 0$. The *Brownian motion with drift* defined by

  $$X_t = \mu t + \sigma B_t$$

  is a Lévy process. Indeed, the Brownian motion with linear drift is in the only Lévy process with continuous sample paths.
- A jump-diffusion process such as

  $$X_t = \mu t + \sigma B_t + J_t \ ,$$

  where B_t is a standard Brownian motion and J_t is a compound Poisson process.

3.5 Markov Processes

A Markov process is a stochastic process with the property that the probability of a future event conditioned on all the process past history is equal to the probability of that same future event conditioned only on the present state of the process, i.e. given the present state of the system, its future and past are independent. Examples of Markov processes include the Poisson process and Brownian motion.

The formal definition of a Markov process is the following.

Definition 3.5.1 (Markov process)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space with a filtration $\{\mathcal{F}_t\}$ and let μ be a probability measure on $\mathcal{B}(\mathbb{R}^d)$. An adapted stochastic process $X = \{X_t : t \in [0, +\infty)\}$ with values in $\mathbb{R}^d$ is called a* Markov process *with initial distribution μ if:*

1) $P(X_0 \in A) = \mu(A)$ for any $A \in \mathcal{B}(\mathbb{R}^d)$.

2) If $s, t > 0$ and $A \in \mathcal{B}(\mathbb{R}^d)$, then

$$P(X_{s+t} \in A | \mathcal{F}_s) = P(X_{s+t} \in A | X_s) \qquad \textit{almost surely.}$$

It is also useful to introduce the concept of a Markov family.

Definition 3.5.2 (Markov family)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space with a filtration $\{\mathcal{F}_t\}$, μ be a probability measure on $\mathcal{B}(\mathbb{R}^d)$ and $X^x = \{X_t^x : t \in [0, +\infty)\}$, $x \in \mathbb{R}^d$, be a family of processes with values in $\mathbb{R}^d$ which are adapted to the filtration $\{\mathcal{F}_t\}$. This family of processes is called a time-homogeneous Markov family if:*

1) The function $p : \mathbb{R}_0^+ \times \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}$ defined by

$$p(t, x, A) = P(X_t^x \in A)$$

is Borel-measurable as a function of $x \in \mathbb{R}^d$ for any $t > 0$ and any set $A \in \mathcal{B}(\mathbb{R}^d)$.

2) $P(X_0^x = x) = 1$ for any $x \in \mathbb{R}^d$.

3) If $s, t > 0$, $x \in \mathbb{R}^d$, and $A \in \mathcal{B}(\mathbb{R}^d)$, then

$$P(X_{s+t}^x \in A | \mathcal{F}_s) = p(t, X_s^x, A) \qquad \textit{almost surely.}$$

The function $p(t, x, A)$ introduced in item 1) of the previous definition is called the *transition function* for the Markov family X_t^x. It has the following properties:

- For fixed $t \geq 0$ and $x \in \mathbb{R}^d$, $p(t, x, A)$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$.
- For fixed $t \geq 0$ and $A \in \mathcal{B}(\mathbb{R}^d)$, $p(t, x, A)$ is a measurable function of $x \in \mathbb{R}^d$.
- $p(0, x, x) = 1$ for every $x \in \mathbb{R}^d$.
- If $s, t \geq 0$, $x \in \mathbb{R}^d$, and $A \in \mathcal{B}(\mathbb{R}^d)$, then

$$p(s + t, x, A) = \int_{\mathbb{R}^d} p(t, y, A) p(s, x, dy) \ .$$

For an example of a Markov family, consider the process $X_t^x = x + B_t$, where $x \in \mathbb{R}^d$ and B_t is a standard Brownian motion in $\mathbb{R}^d$.

A special case of Markov processes arises when the stochastic process takes values on a discrete set of states. This particular class of Markov processes is known as Markov chains. We will discuss the case of continuous-time Markov chains, before considering the case of (discrete-time) Markov chains.

Definition 3.5.3 (Continuous-time Markov chain)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space with a filtration $\{\mathcal{F}_t\}$. An adapted stochastic process $X = \{X_t : t \in [0, +\infty)\}$ with values in a discrete set $S = \{x_1, x_2, x_3, \ldots\}$ is called a* continuous-time Markov chain *if for every $s, t \geq 0$ and states $i, j, x(u) \in S$, with $u < s$, we have*

$$P(X_{s+t} = j \,|\, X_s = i, X_u = x(u), 0 \leq u < s) = P(X_{s+t} = j \,|\, X_s = i) \ .$$

If, in addition, $P(X_{s+t} = j \,|\, X_s = i)$ is independent of s, then the Markov chain is said to have stationary *or* homogeneous *transition probabilities.*

Note that:

- the amount of time a Markov chain spends in a state before making a transition to a different state is exponentially distributed with parameter λ_i.
- the amount of time a Markov chain spends in a state i, and the next state visited, are independent random variables.
- when the Markov chain leaves a state $i \in S$, it will enter state $j \in S$ with some probability p_{ij} such that

$$\sum_{i \neq j} p_{ij} = 1 \ .$$

Example A simple example of a continuous-time Markov chain is provided by *birth and death processes*. Take for state space the set $S = \mathbb{N}_0$ and let q_{ij} the transition rate from state i to state j:

$$q_{ij} = \lambda_i p_{ij} \ ,$$

where λ_i is the rate at which the process leaves state $i \in S$, and p_{ij} is the probability that the process goes to state $j \in S$ from state $i \in S$. A *birth and death process* is a continuous-time Markov chain for which $q_{ij} = 0$ for all $i, j \in S$ such that $|i - j| > 1$. Hence, if the Markov chain is at state i, then it can only go to either state $i - 1$ or state $i + 1$. In applications, the state of the process is usually though of as representing the size of a population. If the state increases by one unit, a birth is said to occur, and if the state decreases by one unit, a death is said to occur. The values

$$b_i = q_{ii+1} \ , \qquad d_i = q_{ii-1}$$

are called, respectively, the birth and death rates. Since $\sum_{i \in \mathbb{N}_0} q_{ij} = \lambda_i$, we obtain that

$$\lambda_i = b_i + d_i \ , \qquad p_{ii+1} = 1 - p_{ii-1} = \frac{b_i}{b_i + d_i} \ .$$

We consider now the discrete-time case of a Markov chain.

Definition 3.5.4 (Discrete-time Markov chain)**.** *A stochastic process $X = \{X_t : t \in \mathbb{N}_0\}$ on a probability space $(\Omega, \mathcal{F}, P)$ with values on a discrete set $S = \{x_1, x_2, x_3, \ldots\}$ is called a* Markov chain *if for every $t \in \mathbb{N}_0$, every $i, j \in S$ and every sequence $i_0, i_1, \ldots, i_{t-1} \in S$, we have*

$$P(X_{t+1} = j \,|\, X_t = i, X_{t-1} = i_{t-1}, \ldots, X_0 = i_0) = P(X_{t+1} = j \,|\, X_t = i) \ .$$

If $P(X_{t+1} = j \,|\, X_t = i)$ is independent of t, the Markov chain is said to have stationary *or* homogeneous *transition probabilities.*

Let X be a Markov chain and let p_{ij} denote the probability that the process X will make a transition to state $j \in S$ from state $i \in S$, i.e.

$$p_{ij} = P(X_{t+1} = j \,|\, X_t = i) \ .$$

Then, we have that

$$p_{ij} \geq 0 \quad \text{for all } i, j \in S \text{ and } \quad \sum_{j \in S} p_{ij} = 1 \ .$$

Example The random walk provides an example of a Markov Chain. Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of independent identically distributed random variables satisfying

$$P(X_i = j) = a_j \ , \qquad j \in \mathbb{Z} \ .$$

Define a stochastic process $S = \{S_n : n \in \mathbb{N}_0\}$ by

$$S_0 = 0 \ , \qquad S_n = \sum_{i=1}^{n} X_i \ .$$

Then, S is a Markov chain with transition probabilities given by $p_{ij} = a_{j-i}$.

4 Statistics

This last section is devoted to an overview of some topics in Statistics. We will provide a very brief review of some concepts in *statistical inference*, including random sampling, estimation, confidence interval and hypothesis testing.

4.1 Random sample and Statistic

The act of collecting data through observation of a variable of interest in some experiment is very often the first step in the statistical treatment of such experiment. We will refer to the set of all units of a given class (under observation) as a population. That class may be people, buildings, physical quantities or economical results. The term population is also commonly used to refer to the set of all potential measurements or values in a given experiment.

Clearly, a population may be of finite or infinite size. Even when the population under consideration is finite, it is usually not practical to observe every one of its elements. The typical procedure to follow is to extract a subset of the population of appropriate size. Such process is called sampling and the resulting subset is called sample. Depending on the population under observation, different types of sampling may be used. For instance, one may wish to divide the full population into smaller subsets of elements sharing some particular property, and only then sample from each of these sets. This process is known as cluster sampling or stratified sampling.

In what follows, we will only consider random sampling. By random sampling one may either refer to a set of independent and identically distributed random variables corresponding to some given observation or measurement, or to a sample of individuals selected from a population in such a way that each sample of the same size is equally likely. We will use the former nomenclature here.

Definition 4.1.1 (Random sample)**.** *The random variables $X_1, \dots X_n$ are called a* random sample of size n from the population $f(x)$ *if $X_1, \dots, X_n$ are mutually independent and identically distributed random variables with probability function $f(x)$ (density in the continuous case).*

A random sample corresponds to an experimental situation in which the variable of interest has a probability distribution described by $f(x)$. In most experiments there are $n > 1$ repeated observations made on the variable, the first being X_1, the second X_2, and so on. Note that each X_i is an observation on the same variable and each X_i has distribution determined by $f(x)$. Moreover, the observations are made in such a way that the value of one observation is independent of any of the other observations. Thus, we obtain that the joint probability function $f(x_1, \dots, x_n)$ (density in the continuous case) of the random sample $X_1, \dots, X_n$ is such that

$$f(x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i) \ .$$

If the population under observation is assumed to have a specified parametric family of probability distributions $f(x|\theta)$ with unknown true parameter value $\theta \in \mathbb{R}^K$, then a random sample extracted from this population has a joint probability

function (density in the continuous case) $f(x_1, \ldots, x_n|\theta)$ satisfying:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \ .$$

By considering different values for the parameter θ, we can study how a random sample behaves for different populations. On the other hand, we can use the random sample to estimate the value of the parameter θ and move forward with the statistical analysis from this point.

When a sample $X_l, \ldots, X_n$ is extracted from a population, one may try to construct relevant quantities describing the main properties of observed sample. Such quantities may be expressed in the form of a function $T(x_1, \ldots, x_n)$, which may be real-valued or vector-valued, and whose domain includes the sample space of the random vector $(X_1, \ldots, X_n)$. Thus, these quantities define a random variable (or vector) $Y = T(X_1, \ldots, X_n)$ known as a statistic.

Definition 4.1.2 (Statistic)**.** *Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T : \mathbb{R}^n \to \mathbb{R}^k$, $k \geq 1$, be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. The random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a* statistic *and its probability distribution is called the* sampling distribution of Y*.*

Note that a statistic is a function of the random sampling only, and does not involve any other unknown parameters. Examples of a statistic include the sample mean and the sample variance defined below.

Definition 4.1.3 (Sample mean)**.** *Let $X_1, \ldots, X_n$ be a random sample of size n. The* sample mean*, denoted by $\overline{X}_n$, is the arithmetic average of the values in the random sample:*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \ .$$

Definition 4.1.4 (Sample variance)**.** *Let $X_1, \ldots, X_n$ be a random sample of size n. The* sample variance*, denoted by S^2, is the statistic defined by*

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \ .$$

The next result lists some properties of the sample mean and variance.

Proposition 4.1.5 (Properties of the sample mean and variance)**.** *Let $X_1, \ldots, X_n$ be a random sample from a population with finite mean μ and variance σ^2. Then:*

1) $E[\overline{X}_n] = \mu$.

2) $\mathrm{Var}[\overline{X}_n] = \sigma^2/n$.

3) $E[S^2] = (n-1)\sigma^2/n$.

Due to the fact that $E[S^2] \neq \sigma^2$, a modified version of the sample variance is more commonly used.

Definition 4.1.6 (Corrected sample variance)**.** *Let $X_1, \ldots, X_n$ be a random sample of size n. The* corrected sample variance*, denoted by S'^2, is the statistic defined by*

$$S'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \ .$$

Let $X_1, \ldots, X_n$ be a random sample from a population with finite mean μ and variance σ^2. It is possible to check that the mathematical expectation of the corrected sample variance is equal to:

$$E[S'^2] = \sigma^2 \ .$$

We now state another very interesting property of S'^2, widely used in statistical inference. Let $X_1, \ldots, X_n$ be a random sample of size n taken from a normal population $N(\mu, \sigma^2)$. Then, the random variable

$$Q = \frac{(n-1)S'^2}{\sigma^2} \sim \chi^2(n-1) \ .$$

It is very common that an explicit expression for the exact distribution of a statistic $Y = T(X_1, \ldots, X_n)$

$$F_Y(y) = P((x_1, \ldots, x_n) \in \mathbb{R}^n : T(x_1, \ldots, x_n) \leq y)$$

is not available. To avoid this sort of difficulty, one may resort to asymptotic distributions (such as the one provided by the Central Limit Theorem) or to the Monte Carlo method when there is no analytical description available for the statistic distribution.

One particular class of statistics for which it is possible to compute its exact distribution are *order statistics*. Let $X_1, \ldots, X_n$ be a random sample of size n with distribution function $F(x)$ and probability function (density in the continuous case) $f(x)$. The order statistics of $X_1, \ldots, X_n$ are the values obtained by ordering the random sample:

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \ .$$

Hence, the value $X_{(1)}$ corresponds to the sample minimum

$$X_{(1)} = \min\{X_1, \ldots, X_n\}$$

and the value $X_{(n)}$ corresponds to the sample maximum

$$X_{(n)} = \max\{X_1, \ldots, X_n\} \ .$$

The distribution of the order statistic $X_{(i)}$, $i \in \{1, \ldots, n\}$, is determined by its probability function (or density) $f_{X_{(i)}}$ given by

$$f_{X_{(i)}}(x) \quad = \quad \frac{n!}{(i-1)!(n-i)!} \left(F(x)\right)^{i-1} \left(1 - F(x)\right)^{n-i} f(x) \ .$$

4.2 Estimators

Consider the case where a random sample is taken from a population with probability function (or density) $f(x|\theta)$ depending on an unknown parameter θ. Any knowledge about the parameter θ leads to knowledge about the entire population. Thus, methods providing a good estimator for the value of θ are of great relevance for Statistics. Very often, the parameter θ has also a meaningful interpretation such as, for instance, the case of a population mean.

Definition 4.2.1 (Point estimator)**.** *Let $X_1, \ldots, X_n$ be a random sample of size n taken from a population with probability function (density in the continuous case) $f(x|\theta)$, where $\theta \in \Theta$ is an unknown parameter. A* point estimator for θ *is a statistic $\hat{\theta}(X_1, \ldots, X_n)$ that is used to infer the value of θ.*

Note the distinction between an estimator and an estimate. An estimator is a function of the random sample alone, while an estimate is the numerical value of an estimator obtained when a sample is actually taken.

There are several techniques that can be used to construct estimators. We will discuss two of these techniques here – the *Method of Moments* and the *Maximum Likelihood Estimator*. We discuss the former method first.

4.2.1 Method of Moments

Let $X_1, \ldots, X_n$ be a random sample of size n taken from a population with probability function (density in the continuous case) $f(x|\theta_1, \ldots, \theta_k)$, $k \geq 1$. Assume that the distribution determined by $f(x|\theta_1, \ldots, \theta_k)$ has as many moments $E[X^r]$ as needed. Let $m_1, \ldots, m_k$ denote the sample moments

$$m_j = \frac{1}{n} \sum_{i=1}^{n} X_i \ , \qquad j \in \{1, \ldots, k\} \ ,$$

readily computable from the sample, and let $\mu'_1, \ldots, \mu'_k$ denote the corresponding population moments

$$\mu'_j = E[X^j] \ , \qquad j \in \{1, \ldots, k\} \ .$$

Note that the population moments are functions of the parameters $\theta_1, \ldots, \theta_k$, i.e. $\mu'_j = \mu'_j(\theta_1, \ldots, \theta_k)$. The method of moments estimators are then found by solving with respect to $\theta_1, \ldots, \theta_k$ the system of k equations in k unknowns given by

$$\begin{aligned} \mu'_1(\theta_1, \ldots, \theta_k) &= m_1 \\ &\vdots \\ \mu'_k(\theta_1, \ldots, \theta_k) &= m_k \ . \end{aligned}$$

If this system turns out to be underdetermined, introduce more equations using higher order moments.

Example For an example of application of the method of moments, let $X_1, \ldots, X_n$ be a random sample of size n taken from an exponential population with unknown parameter $\lambda > 0$. Recalling that the expected variable of a random variable $X \sim$

$Ex(\lambda)$ is $E[X] = 1/\lambda$, we obtain that the method of moments estimator is the solution of the following equation (with respect to λ):

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i \ ,$$

Thus, we obtain the estimator

$$\hat{\lambda}(X_1, \ldots, X_n) = \frac{n}{\sum_{i=1}^{n} X_i}$$

for the unknown parameter λ.

4.2.2 Maximum Likelihood method

We will now introduce one of the most used methods to obtain estimators – the maximum likelihood method. As before, let $X_1, \ldots, X_n$ be a random sample taken from a population with probability function (or density) $f(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}^K$, $K \geq 1$, is an unknown parameter. The *likelihood function* $L : \Theta \to \mathbb{R}$ is defined by

$$L(\theta|x) = L(\theta_1, \ldots, \theta_k \,|\, x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k) \ .$$

A *maximum likelihood estimator* $\hat{\theta}(X_1, \ldots, X_n)$ for θ is the value of θ for which the likelihood function $L(\theta|x)$ attains its maximum as a function of θ, while x is held fixed.

The maximum likelihood estimator is very often a reasonable choice for an estimator. To find the maximum likelihood estimator one has to deal with the problem of finding the global maximum of a function. If the probability function (or density function) of the random sample is reasonably behaved, this problem reduces to a standard calculus problem, even though the computations may be cumbersome in some cases.

If the likelihood function is differentiable with respect to $\theta \in \Theta$, possible candidates for the maximum likelihood estimator are the values of $\theta = (\theta_1, \ldots, \theta_K)$ satisfying the first order conditions

$$\frac{\partial L}{\partial \theta_i}(\theta|x) = 0 \ , \qquad i = 1, \ldots, K \ .$$

Note that any solution of the set of equations above is only a candidate for a maximum likelihood estimator, i.e. the first order condition above is only a necessary condition for a maximum, not a sufficient one. Moreover, the zeros of the first derivative of L locate only extreme points in the interior of the set Θ. If the extrema occur on the boundary of Θ, the first derivative may not be zero. Thus, the boundary must be checked separately for extrema.

Example Let $X_1, \ldots, X_n$ be a random sample of size n taken from a normal pop-

ulation $N(\mu, 1)$ with unknown mean μ. The likelihood function $L : \mathbb{R} \to \mathbb{R}$ is

$$\begin{aligned} L(\mu|x) &= \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \mathrm{e}^{-(x_i-\mu)^2/2} \\ &= \frac{1}{(2\pi)^{n/2}} \mathrm{e}^{-(1/2)\sum_{i=1}^{n}(x_i-\mu)^2} \ . \end{aligned}$$

Working out the equation

$$\frac{\partial L}{\partial \mu}(\mu|x_1, \ldots, x_n) = 0$$

we obtain

$$\sum_{i=1}^{n}(x_i - \mu) = 0 \ .$$

Hence, one gets the solution

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \ ,$$

as expected. To check that $\hat{\mu}$ is indeed a maximum likelihood estimator for μ, it is enough to check that

$$\frac{\partial^2 L}{\partial \mu^2}(\hat{\mu}|x_1, \ldots, x_n) < 0 \ ,$$

which turns out to be the case here.

Before concluding the discussion of maximum likelihood estimators, one should point out one special property of this class of estimators – *the invariance principle*. Let $\hat{\theta}$ be the maximum likelihood estimator for the parameter $\theta \in \Theta \subset \mathbb{R}^K$. Then, for any function h of the unknown parameter θ, the maximum likelihood estimator of $h(\theta)$ is given by $h(\hat{\theta})$.

4.2.3 Some measures to assess estimators quality

We will now discuss some desirable properties for an estimator to have. In what follows, let $X_1, \ldots, X_n$ be a random sample of size n taken from a population with probability function (density in the continuous case) $f(x|\theta)$, where $\theta \in \Theta$ is an unknown parameter.

Definition 4.2.2 (Bias)**.** *The* bias *of an estimator $\hat{\theta}$ of a parameter θ, denoted by* $\mathrm{Bias}_\theta(\hat{\theta})$*, is*

$$\mathrm{Bias}_\theta(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta \ ,$$

where E_θ denotes the expected value taken with respect to $f(x|\theta)$. An estimator whose bias is identically zero as a function of θ is said to be unbiased *and satisfies $E_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Theta$.*

Example Let $X_1, \ldots, X_n$ be a random sample from a normal population $N(\mu, \sigma^2)$. We have already seen above that $E_{(\mu,\sigma^2)}[\overline{X}_n] = \mu$, i.e. the sample average $\overline{X}_n$ is an unbiased estimator for the population mean μ (this statement hold for non-normal populations also). In what concerns the population variance, we have seen that $E_{(\mu,\sigma^2)}[S^2] = (n-1)\sigma^2/n$, and thus the sample average has non-zero bias:

$$\text{Bias}_{(\mu,\sigma^2)}(S^2) = -\frac{\sigma^2}{n} \ .$$

This is a disadvantage of using S^2 instead of the corrected sample average S'^2, since S'^2 is an unbiased estimator for σ^2.

Definition 4.2.3 (Mean square error)**.** *The* mean squared error *of an estimator $\hat{\theta}$ of a parameter θ is the function of θ defined by $E_\theta[(\hat{\theta} - \theta)^2]$.*

The mean squared error measures the average squared difference between the estimator $\hat{\theta}$ and the unknown parameter θ. It provides a reasonable measure for the performance of an estimator, even though any increasing function of the distance $|\hat{\theta} - \theta|$ would provide the same sort of information. However, in the case of the mean square error it is possible to prove that the following identity is satisfied:

$$E_\theta[(\hat{\theta} - \theta)^2] = \text{Var}_\theta[\hat{\theta}] + \Big(\text{Bias}_\theta(\hat{\theta})\Big)^2 \ ,$$

where Var_θ denotes the variance taken with respect to the population probability function (or density) $f(x|\theta)$. Hence, the mean square error may be decomposed into two components, one measuring the estimator variability or precision, and the other measuring its bias. Thus, an estimator with small mean square error has small combined variance and bias.

Example Returning to the previous example, where $X_1, \ldots, X_n$ is a random sample from a normal population $N(\mu, \sigma^2)$, one can see that

$$E_{(\mu,\sigma^2)}[(\overline{X}_n - \mu)^2] = \text{Var}_{(\mu,\sigma^2)}(\overline{X}_n) = \frac{\sigma^2}{n} \ .$$

Thus, the mean square error of the sample mean decreases as the sample size increases. In what concerns the sample variance S^2 and the corrected sample variance S'^2, it is possible to check that

$$\begin{aligned} E_{(\mu,\sigma^2)}[(S^2 - \sigma^2)^2] &= \text{Var}_{(\mu,\sigma^2)}[S^2] + \text{Bias}_{(\mu,\sigma^2)}(S^2) \\ &= \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} \\ &= \frac{(2n-1)\sigma^4}{n^2} \end{aligned}$$

while

$$E_{(\mu,\sigma^2)}[(S'^2 - \sigma^2)^2] = \text{Var}_{(\mu,\sigma^2)}[S'^2] = \frac{2\sigma^4}{n-1} \ .$$

Thus, even though S'^2 is unbiased while S^2 is biased, we obtain that

$$E_{(\mu,\sigma^2)}[(S^2 - \sigma^2)^2] < E_{(\mu,\sigma^2)}[(S'^2 - \sigma^2)^2]$$

for every $n \geq 2$.

This example does not imply that S^2 should be used as an estimator of σ^2 instead of S'^2. The computation above only shows that on the average S^2 is closer to σ^2 than S'^2 if the distance is measured using the mean square error. However, S^2 is biased and will underestimate σ^2 on average.

Note that controlling bias does not guarantee that the mean square error is controlled. Indeed, it may even happen that a trade-off occurs between variance and bias in such a way that an increase in bias leads to a larger decrease in variance, resulting in an improvement of an estimator with respect to the mean square error. Moreover, a comparison between two different estimators $\hat{\theta}_1, \hat{\theta}_2$ of θ may not even yield one "best" estimator. The typical situation is that one estimator is better with respect to the other in a subset of the parameter space Θ, while the opposite is true in the complementary of that subset.

In order to use mean square error to compare two estimators, one must restrict the analysis to smaller classes of estimators. One such class is the one whose elements are unbiased estimators. In this case, if $\hat{\theta}_1, \hat{\theta}_2$ are unbiased estimators of a parameter θ, then their mean squared errors are equal to their variances, and one should choose the estimator with smaller variance.

Definition 4.2.4 (Best unbiased estimator)**.** *An estimator θ^* is a* best unbiased estimator *of θ if it satisfies $E_\theta[\theta^*] = \theta$ for all $\theta \in \Theta$ and, for any other estimator $\hat{\theta}$ with $E_\theta[\hat{\theta}] = \theta$, we have $\mathrm{Var}_\theta[\theta^*] \leq \mathrm{Var}_\theta[\hat{\theta}]$ for all $\theta \in \Theta$.*

Finding a best unbiased estimator, if there is one, may not be easy. First, when comparing two estimators, the computations leading to their variances may be lengthy. Moreover, there may exist another estimator with even smaller variance. The Cramér-Rao inequality provides a lower bound for the variance of an estimator. If one is able to find an unbiased estimator with variance equal to such lower bound, than this must be a best unbiased estimator.

Theorem 4.2.5 (Cramér-Rao inequality)**.** *Let $X_1, \ldots, X_n$ be a random sample of size n taken from a population with probability function (density in the continuous case) $f(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}$, and let $\hat{\theta}$ be any estimator of θ satisfying*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta[\hat{\theta}] = \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} \left[\hat{\theta}(x_1, \ldots, x_n) f(x_1, \ldots, x_n|\theta) \right] \mathrm{d}x$$

and

$$\mathrm{Var}_\theta[\hat{\theta}] < \infty \ .$$

Then,

$$\mathrm{Var}_\theta[\hat{\theta}] \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta[\hat{\theta}] \right)^2}{I(\theta)} \ ,$$

where $I(\theta)$ is the Fisher information *of the sample*

$$I(\theta) = E_\theta \left[\left(\frac{\partial}{\partial \theta} \ln f(x_1, \ldots, x_n)|\theta) \right)^2 \right] \ .$$

Note that although the Cramér-Rao inequality is stated above for continuous random variables, it also applies to discrete random variables with the obvious modifications.

If $\hat{\theta}$ is an unbiased estimator for θ, the Cramér-Rao inequality reduces to

$$\mathrm{Var}_\theta[\hat{\theta}] \geq \frac{1}{I(\theta)} \ .$$

It should be remarked that the lower bound provided by Cramér-Rao inequality may not be sharp, i.e. its value may be strictly smaller than the variance of any unbiased estimator. Define an estimator *efficiency* through

$$e(\hat{\theta}) = \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta[\hat{\theta}]\right)^2}{I(\theta)\mathrm{Var}_\theta[\hat{\theta}]} \ .$$

From Cramér-Rao inequality, one obtains that

$$e(\hat{\theta}) \leq 1 \ .$$

An estimator $\hat{\theta}$ is said to be *efficient* if $e(\hat{\theta}) = 1$, i.e. the variance of $\hat{\theta}$ is equal to the lower bound in Cramér-Rao inequality. However, there may be cases where such lower bound is not attainable, i.e. $e(\hat{\theta}) < 1$ for every estimator $\hat{\theta}$ of θ.

4.3 Confidence intervals

As discussed in the previous section, given a random sample $X_1, \ldots, X_n$ taken from a population with probability function (density in the continuous case) $f(x|\theta)$, where $\theta \in \Theta$ is an unknown parameter, one can construct an estimator for θ. This corresponds to finding a function of the random sample yielding a point value for θ. An alternative approach to estimate θ is to look for an entire interval (or region) of plausible values – a *confidence interval*.

When constructing a confidence interval, one needs to select a confidence level, measuring the degree of reliability of the interval. Information about the precision of a confidence interval is then given by the width of the interval. If the confidence level is high and the confidence interval is narrow, the information obtained for the value of the unknown parameter may be rather precise.

Definition 4.3.1 (Confidence interval)**.** *Let $X_1, \ldots, X_n$ be a random sample of size n taken from a population with probability function (density in the continuous case) $f(x|\theta)$, where $\theta \in \Theta$ is an unknown parameter. A* confidence interval for θ with confidence level $1 - \alpha$ *is a random interval*

$$I(X_1, \ldots, X_n) = (L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)) \ ,$$

with the following properties:

i) the functions $L, U : \mathbb{R}^n \to \mathbb{R}$ are statistics such that for every $(x_1, \ldots, x_n) \in \mathbb{R}^n$ the following inequality holds

$$L(x_1, \ldots, x_n) \leq U(x_1, \ldots, x_n) \ .$$

ii) the confidence coefficient of $I(X_1, \ldots, X_n)$ *is* $1 - \alpha$*, i.e.*

$$\inf_{\theta \in \Theta} P_\theta \left[L(X_1, \ldots, X_n) < \theta < U(X_1, \ldots, X_n) \right] = 1 - \alpha \ ,$$

where P_θ *denotes the probability measure determined by* $f(x|\theta)$*.*

In a similar way, we define a *lower confidence bound for* θ *with confidence level* $1 - \alpha$ to be an interval of the form

$$I(X_1, \ldots, X_n) = (L(X_1, \ldots, X_n), +\infty)$$

and an *upper confidence bound for* θ *with confidence level* $1 - \alpha$ to be an interval of the form

$$I(X_1, \ldots, X_n) = (-\infty, U(X_1, \ldots, X_n)) \ .$$

All these notions can be easily extended to define confidence regions for multidimensional parameters θ. However, for simplicity of exposition, we restrict our attention to the one-dimensional case.

We will now describe a general strategy that may be used to construct confidence intervals. Let $X_1, X_2, \ldots, X_n$ be a random sample taken from a population with probability function (density in the continuous case) $f(x|\theta)$, where $\theta \in \Theta$ is the unknown parameter to be estimated.

The first step is to find a random variable $Q : \mathbb{R}^n \times \Theta \to \mathbb{R}$, depending only on the random sample $X_1, \ldots, X_n$ and on the parameter θ, such that the probability distribution of Q does not depend on θ or on any other unknown parameter. The random variable Q is called a *pivot* or a *pivotal quantity*. The precise form of the pivot Q is usually obtained by examining the distribution of an appropriate estimator $\hat{\theta}$ for θ.

Assume now that a pivot Q is provided. Then, for any α such that $0 < \alpha < 1$, there exist constants $a, b \in \mathbb{R}$ such that

$$P_\theta \left(a < Q(X_1, \ldots, X_n, \theta) < b \right) = 1 - \alpha$$

If the inequalities $a < Q(X_1, \ldots, X_n, \theta) < b$ can be worked out in such a way that θ becomes isolated, one obtains the equivalent statement

$$P_\theta \left(L(X_1, \ldots, X_n) < \theta < U(X_1, \ldots, X_n) \right) = 1 - \alpha \ ,$$

where $L(X_1, \ldots, X_n)$ and $U(X_1, \ldots, X_n)$ are, respectively, the lower and upper ends of a confidence interval for θ with confidence level $1 - \alpha$.

A final remark concerning the choice of constants $a, b \in \mathbb{R}$ in the construction above is in order. It should be clear that there are infinitely many choices for these constants preserving the confidence level $1 - \alpha$. Out of all the possible choices, we could ask for the values of a and b minimizing the length $b - a$ of the interval $[a, b]$. The following result provides the appropriate choice for the case where the pivot Q is a continuous random variable with a unimodal probability density function $f(x)$, i.e. there exists $x^* \in \mathbb{R}$ such that $f(x)$ is nondecreasing for $x < x^*$ and $f(x)$ is nonincreasing for $x > x^*$. Among others, this property is shared by the density functions of the following families of probability distributions: normal, chi-square, Student's t, and Snedecor's F.

Proposition 4.3.2. *Let $f(x)$ be a unimodal probability density function. If the interval $[a, b]$ satisfies:*

i) $\int_a^b f(x)\mathrm{d}x = 1 - \alpha$,

ii) $f(a) = f(b) > 0$, *and*

iii) $a \leq x^* \leq b$, *where x^* is a mode of $f(x)$,*

then $[a, b]$ is the shortest among all intervals that satisfy property (i).

We will now discuss several relevant examples of confidence intervals for parameters of normal populations and for large samples.

4.3.1 Mean

We distinguish between the following three cases:

1) Let $X_1, \ldots, X_n$ be a random sample of size n taken from a normal population $N(\mu, \sigma^2)$ with unknown mean μ and known variance σ^2. We have seen above that the sample mean $\overline{X}_n$ is an estimator for the population mean μ with distribution

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) .$$

Thus, the random variable

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

is a pivot for the population mean μ. The next step is to choose $a, b \in \mathbb{R}$ such that

$$P\left(a < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < b\right) = 1 - \alpha .$$

The choice $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$, where $z_{\alpha/2} \in \mathbb{R}$ is the unique solution of $P(Z > z_{\alpha/2}) = \alpha/2$, yields the following confidence interval for μ:

$$\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) .$$

2) Let $X_1, \ldots, X_n$ be a random sample of size n taken from a normal population $N(\mu, \sigma^2)$ with unknown mean μ and variance σ^2. Since σ^2 is unknown, the function

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

is no longer a pivot for the population mean μ. To obtain a pivot we proceed as follows. Recall that

$$U = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

and that

$$V = \frac{(n-1)S'^2}{\sigma^2} \sim \chi^2(n-1) .$$

We obtain that

$$T = \frac{U}{\sqrt{V/(n-1)}} = \frac{\overline{X}_n - \mu}{S'/\sqrt{n}} \sim t(n-1)$$

is a pivot for the population mean μ. Picking values $a = -t_{\alpha/2,n-1}$ and $b = t_{\alpha/2,n-1}$, where $t_{\alpha/2,n-1} \in \mathbb{R}$ is the unique solution of $P(T > t_{\alpha/2,n-1}) = \alpha/2$, yields

$$P\left(a < \frac{\overline{X}_n - \mu}{S'/\sqrt{n}} < b\right) = 1 - \alpha \ .$$

We obtain the following confidence interval for μ:

$$\left(\overline{X}_n - t_{\alpha/2,n-1}\frac{S'}{\sqrt{n}}, \overline{X}_n + t_{\alpha/2,n-1}\frac{S'}{\sqrt{n}}\right) \ .$$

3) Let $X_1, \ldots, X_n$ be a random sample of large size n taken from a population with unknown mean μ and variance σ^2. In this case, even if an appropriate pivotal quantity is not available, provided the sample size n is sufficiently large, the central limit theorem may be used to obtain an *approximate confidence interval* for the mean μ of the population. In such case, the pivotal quantity is

$$Z = \frac{\overline{X}_n - \mu}{S'/\sqrt{n}} \overset{a}{\sim} N(0,1) \ ,$$

where the notation $\overset{a}{\sim}$ is used to denote that Z is asymptotically normally distributed when $n \to \infty$. Proceeding similarly to case 1) above, one obtains the following approximate confidence interval for μ:

$$\left(\overline{X}_n - z_{\alpha/2}\frac{S'}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{S'}{\sqrt{n}}\right) \ .$$

If n is not very large, one can use the Student's t distribution of case 2), yielding a slightly wider approximate confidence interval.

4.3.2 Variance

Let $X_1, \ldots, X_n$ be a random sample of size n taken from a normal population $N(\mu, \sigma^2)$ with unknown mean μ and variance σ^2. The pivotal quantity to be used is

$$Q = \frac{(n-1)S'^2}{\sigma^2} \sim \chi^2(n-1) \ .$$

Let $q_{\alpha,n-1}$ be such that $P(Q > q_{\alpha,n-1}) = \alpha$. Picking values $a = q_{1-\alpha/2,n-1}$ and $b = q_{\alpha/2,n-1}$ yields

$$P\left(a < \frac{(n-1)S'^2}{\sigma^2} < b\right) = 1 - \alpha \ .$$

We obtain the following confidence interval for σ^2:

$$\left(\frac{(n-1)S'^2}{q_{\alpha/2,n-1}}, \frac{(n-1)S'^2}{q_{1-\alpha/2,n-1}}\right) \ .$$

Note that since the chi-square distribution is non-symmetric, the choice made above for the values a, b does not yield the confidence interval with confidence level $1 - \alpha$ with smaller expected width.

4.3.3 Difference of two means

We distinguish between the following four cases:

1) Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken, respectively, from two normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with unknown means μ_1 and μ_2 and known variances σ_1^2 and σ_2^2. Using the properties of the normal distribution, it is possible to check that

$$Z = \frac{(\overline{X_{1n}} - \overline{X_{2m}}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1) \ ,$$

where $\overline{X_{1n}}, \overline{X_{2m}}$ denote the sample means of each random sample. Thus, Z is a pivotal quantity. Picking values $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$, where $z_{\alpha/2} \in \mathbb{R}$ is the unique solution of $P(Z > z_{\alpha/2}) = \alpha/2$, yields

$$P\left(a < Z < b\right) = 1 - \alpha \ .$$

We obtain the following confidence interval for $\mu_1 - \mu_2$:

$$\left(\overline{X_{1n}} - \overline{X_{2m}} - z_{\alpha/2}\sigma^*, \overline{X_{1n}} - \overline{X_{2m}} + z_{\alpha/2}\sigma^*\right) \ ,$$

where

$$\sigma^* = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \ .$$

2) Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken, respectively, from two normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with unknown means μ_1 and μ_2 and unknown, but equal, variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$. It is possible to check that

$$T = \frac{(\overline{X_{1n}} - \overline{X_{2m}}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_1'^2 + (m-1)S_2'^2}{n+m-2}}} \sim t(n+m-2)$$

is a pivotal quantity for the difference $\mu_1 - \mu_2$. Picking values $a = -t_{\alpha/2,n+m-2}$ and $b = t_{\alpha/2,n+m-2}$, where $t_{\alpha/2,n+m-2} \in \mathbb{R}$ is the unique solution of $P(T > t_{\alpha/2,n+m-2}) = \alpha/2$, yields

$$P\left(a < T < b\right) = 1 - \alpha \ .$$

We obtain the following confidence interval for $\mu_1 - \mu_2$:

$$\left(\overline{X_{1n}} - \overline{X_{2m}} - t_{\alpha/2,n+m-2}S^*, \overline{X_{1n}} - \overline{X_{2m}} + t_{\alpha/2,n+m-2}S^*\right) \ ,$$

where

$$S^* = \sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_1'^2 + (m-1)S_2'^2}{n+m-2}} \ .$$

3) Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken, respectively, from two normal population $N(\mu_1, \sigma_1^2)$ and

$N(\mu_2, \sigma_2^2)$ with unknown means μ_1 and μ_2 and unknown and unequal variances σ_1^2 and σ_2^2. It is possible to check that the following asymptotic identity holds:

$$T = \frac{(\overline{X_{1n}} - \overline{X_{2m}}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1'^2}{n} + \frac{S_2'^2}{m}}} \overset{a}{\sim} t(r) \ ,$$

where r is the integer part of r^*:

$$r^* = \frac{\left(\frac{S_1'^2}{n} + \frac{S_2'^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_1'^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_2'^2}{m}\right)^2} \ .$$

Thus, T is a pivotal quantity for the difference $\mu_1 - \mu_2$. Picking values $a = -t_{\alpha/2,r}$ and $b = t_{\alpha/2,r}$, where $t_{\alpha/2,r} \in \mathbb{R}$ is the unique solution of $P(T > t_{\alpha/2,r}) = \alpha/2$, yields

$$P\left(a < T < b\right) = 1 - \alpha \ .$$

We obtain the following approximate confidence interval for $\mu_1 - \mu_2$:

$$\left(\overline{X_{1n}} - \overline{X_{2m}} - t_{\alpha/2,r}S^*, \overline{X_{1n}} - \overline{X_{2m}} + t_{\alpha/2,r}S^*\right) \ ,$$

where

$$S^* = \sqrt{\frac{S_1'^2}{n} + \frac{S_2'^2}{m}} \ .$$

4) Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken from two populations with unknown means μ_1 and μ_2 and unknown variances σ_1^2 and σ_2^2. Even if an appropriate pivotal quantity is not available, provided the sample sizes n and m are sufficiently large, the central limit theorem may be used to obtain an *approximate confidence interval* for the difference $\mu_1 - \mu_2$. It is possible to check that

$$Z = \frac{(\overline{X_{1n}} - \overline{X_{2m}}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1'^2}{n} + \frac{S_2'^2}{m}}} \overset{a}{\sim} N(0,1)$$

may be used as a pivotal quantity. Proceeding similarly to case 1) above, one obtains the following confidence interval for $\mu_1 - \mu_2$:

$$\left(\overline{X_{1n}} - \overline{X_{2m}} - z_{\alpha/2}S^*, \overline{X_{1n}} - \overline{X_{2m}} + z_{\alpha/2}S^*\right) \ ,$$

where

$$S^* = \sqrt{\frac{S_1'^2}{n} + \frac{S_2'^2}{m}} \ .$$

4.3.4 Ratio of two variances

Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken, respectively, from two normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with unknown means and variances. To construct a confidence interval for the ratio of the two variances, recall that

$$U = \frac{(n-1)S_1'^2}{\sigma_1^2} \sim \chi^2(n-1)$$

and

$$V = \frac{(m-1)S_2'^2}{\sigma_2^2} \sim \chi^2(m-1) \ .$$

We obtain that

$$F = \frac{U/(n-1)}{V/(m-1)} = \frac{S_1'^2}{S_2'^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(n-1, m-1)$$

is a pivotal quantity for the variances ratio σ_2^2/σ_1^2. Let $f_{\alpha,n-1,m-1}$ be such that $P(F > f_{\alpha,n-1,m-1}) = \alpha$. Picking values $a = f_{1-\alpha/2,n-1,m-1}$ and $b = f_{\alpha/2,n-1,m-1}$ yields

$$P\left(a < \frac{S_1'^2}{S_2'^2}\frac{\sigma_2^2}{\sigma_1^2} < b\right) = 1 - \alpha \ .$$

We obtain the following confidence interval for the ratio σ_2^2/σ_1^2:

$$\left(\frac{S_2'^2}{S_1'^2} f_{1-\alpha/2,n-1,m-1}, \frac{S_2'^2}{S_1'^2} f_{\alpha/2,n-1,m-1}\right) \ .$$

Similarly to the chi-square distribution, Snedecor's F distribution is non-symmetric. Thus, the choice made above for the values a, b does not yield the confidence interval with confidence level $1 - \alpha$ with smaller expected width.

4.3.5 Proportion

Let $X_1, \ldots, X_n$ be a random sample of large size n taken from a Bernoulli population $Bi(1, p)$ with unknown proportion p. It is known that the sample mean $\overline{X}_n$ is a good estimator for p. If the sample size n is sufficiently large, then the central limit theorem may be used to obtain an *approximate confidence interval* for the proportion p. Recall that

$$Z = \frac{\overline{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{a}{\sim} N(0,1) \ .$$

Taking $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$, where $z_{\alpha/2} \in \mathbb{R}$ is the unique solution of $P(Z > z_{\alpha/2}) = \alpha/2$, yields the following approximation

$$P\left(\overline{X}_n - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \overline{X}_n + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha \ .$$

Using $\overline{X}_n$ as an estimate for p, we obtain the following approximate confidence interval for p:

$$\left(\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right) \ .$$

4.3.6 Difference of two proportions

Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken, respectively, from two Bernoulli populations $Bi(1, p_1)$ and $Bi(1, p_2)$ with unknown proportions p_1 and p_2. If the sample sizes n and m are sufficiently large, the central limit theorem may be used to obtain an *approximate confidence interval* for the difference $p_1 - p_2$. It is possible to check that

$$Z = \frac{(\overline{X_{1}}_n - \overline{X_{2}}_m) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}} \overset{a}{\sim} N(0,1) \ .$$

Using $\overline{X_{1}}_n$ as an estimate for p_1 and $\overline{X_{2}}_n$ as an estimate for p_2, and taking $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$, where $z_{\alpha/2} \in \mathbb{R}$ is the unique solution of $P(Z > z_{\alpha/2}) = \alpha/2$, yields the following approximate confidence interval for $\mu_1 - \mu_2$:

$$\left(\overline{X_{1}}_n - \overline{X_{2}}_m - z_{\alpha/2}S^*, \overline{X_{1}}_n - \overline{X_{2}}_m + z_{\alpha/2}S^*\right) \ ,$$

where

$$S^* = \sqrt{\frac{\overline{X_{1}}_n(1 - \overline{X_{1}}_n)}{n} + \frac{\overline{X_{2}}_m(1 - \overline{X_{2}}_m)}{m}} \ .$$

4.4 Hypothesis Testing

The previous two sections were devoted to the topic of estimation. The estimate may be given by a single value or an interval with some given confidence level. Very often, the goal is not to estimate a parameter, but to decide between two contradictory claims about a given parameter. This latter goal is accomplished by the part of statistical inference called *hypothesis testing*.

A *hypothesis* is a statement about a population parameter and the goal of a *hypothesis test* is to decide, based on a random sample taken from the population, which of two alternative hypotheses is true. These two hypotheses are called the *null hypothesis*, denoted by H_0, and the *alternative hypothesis*, denoted by H_1.

If $\theta \in \Theta$ denotes a population parameter, the general format of the null and alternative hypotheses is $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where Θ_0, Θ_1 are subsets of the parameter space Θ such that $\Theta_0 \cap \Theta_1 = \emptyset$. Thus, in a hypothesis testing problem, after observing the sample, one must decide either to reject the null hypotheses H_0 as false or to reject the alternative hypothesis H_1 as false.

Definition 4.4.1 (Hypothesis test)**.** *A* hypothesis test *is a rule that specifies:*

1) For which sample values the null hypothesis H_0 is rejected.

2) For which sample values the alternative hypothesis H_1 is rejected.

The subset of the sample space for which H_0 will be rejected is called the rejection region*, while its complement is called the* acceptance region*.*

For simplicity of exposition, we will only consider null hypotheses of the form $H_0 : \theta = \theta_0$, where $\theta_0 \in \Theta$ is fixed, i.e. $\Theta_0 = \{\theta_0\}$ is a singleton. Alternatives to a null hypothesis of this form include the following three assertions:

i) $H_1 : \theta \neq \theta_0$;

ii) $H_1 : \theta > \theta_0$ (in which case the implicit null hypothesis may be seen as $\theta \leq \theta_0$);

iii) $H_1 : \theta < \theta_0$ (in which case the implicit null hypothesis may be seen as $\theta \geq \theta_0$).

A hypothesis test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ might incur in the following two types of errors. If $\theta \in \Theta_0$ but the hypothesis test result is to reject H_0, then the test has made a *Type I Error*. If, on the other hand $\theta \in \Theta_1$ but the test result is to accept H_0, a *Type II Error* has been made. Ideally, we would like to have test procedures with none of these errors. However, for a fixed sample size, such goal is usually impossible to achieve. The usual strategy to obtain a good test is to restrict ourselves to the class of tests with a prespecified probability of Type I error. Within this class of tests, one can then look for tests with Type II Error probability as small as possible.

Hence, the probabilities of occurrence of each of the two types of errors described above are key parameters in hypothesis testing. These are related with *significance* and *power* of a test. The *significance level* of a test is equal to $1 - \alpha$, where α is the probability of Type I Error:

$$\alpha = P(\text{Type I Error}) = P(\text{reject } H_0 | \theta \in \Theta_0) \ .$$

If Θ_0 has more than one element, then α is the supremum over $\theta \in \Theta_0$ of the probability of Type I Error. The power of a test is equal to $1 - \beta$, where β is probability of Type II Error, i.e.

$$\beta = P(\text{Type II Error}) = P(\text{accept } H_0 | \theta \notin \Theta_0) \ .$$

Therefore, if two distinct statistical tests are available with the same hypotheses H_0 and H_1 and the same level of significance, one can compare their powers and choose the one with larger value.

There are several methods that can be used to construct hypothesis tests. Due to lack of time and space, we skip the precise discussion of any of these methods, but the interested reader can find that information in the references. We restrict our attention to one particular strategy relating the formulation of hypothesis tests with the construction of confidence intervals. This approach is suitable to test the parameters of normal populations, or asymptotically normal quantities, for instance. Its main steps and assumptions may be roughly stated as follows:

1) Consider the hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta \in \Theta_1$, where $\theta_0 \notin \Theta_1$. Assume that for each $\theta_0 \in \Theta$ there is a set $\Theta_1(\theta_0) \subset \Theta$, where $\theta_0 \notin \Theta_1$. Note that this is the case of the alternative hypothesis $H_1 : \theta \neq \theta_0$, $H_1 : \theta > \theta_0$, and $H_1 : \theta < \theta_0$.

2) Depending on the parameter θ and the probability distribution $f(x|\theta)$ of the population from where the random sample is to be extracted, look for an estimator $\hat{\theta}(X_1, \ldots, X_n)$ for θ.

3) Fix the level of significance $1 - \alpha$ at which the test will be performed.

4) Use α and $\hat{\theta}$ to determine the rejection region of the test, i.e. find an appropriate subset $R \subset \mathbb{R}$ such that

$$P(\hat{\theta}(X_1, \ldots, X_n) \in R \,|\, \theta = \theta_0) = \alpha \ .$$

If the distribution of $\hat{\theta}(X_1, \ldots, X_n)$ depends on unknown parameters, one may need to find an appropriate pivotal quantity $Q(X_1, \ldots, X_n, \theta_0)$ and then a rejection region R' such that

$$P(Q(X_1, \ldots, X_n, \theta_0) \in R') = \alpha \ .$$

5) Collect the random sample and reject the null hypothesis or the alternative hypothesis depending on the observed value for $\hat{\theta}$ or $Q(X_1, \ldots, X_n, \theta_0)$.

There is a missing point in the strategy described above: how do we obtain an appropriate rejection region? The answer is provided by the next theorem and explores the close relation between the notions of confidence set and hypothesis test.

Theorem 4.4.2. *For each $\theta_0 \in \Theta$, consider a statistical test with significance level $1 - \alpha$ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1(\theta_0)$ and denote by $\Omega_\alpha(\theta_0)$ its acceptance region, i.e. the subset of sample space Ω for which H_0 is not rejected. For each $(x_1, \ldots, x_n) \in \Omega$, define a set $\Theta_\alpha(x_1, \ldots, x_n)$ in parameter space by*

$$\Theta_\alpha(x_1, \ldots, x_n) = \{\theta \in \Theta : (x_1, \ldots, x_n) \in \Omega_\alpha(\theta)\} \ .$$

Then $\Theta_\alpha(x_1, \ldots, x_n)$ is a $1 - \alpha$ confidence set for θ.
Conversely, let $\Theta_\alpha(x_1, \ldots, x_n)$ be a $1 - \alpha$ confidence set for θ. For any $\theta_0 \in \Theta$, define

$$\Omega_\alpha(\theta_0) = \{(x_1, \ldots, x_n) \in \Omega : \theta_0 \in \Theta_\alpha(x_1, \ldots, x_n)\} \ .$$

Then, $\Omega_\alpha(\theta_0)$ is the acceptance region of a statistical test with significance level $1-\alpha$ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1(\theta_0)$.

Summarizing, the hypothesis test fixes the parameter and asks what sample values, i.e. which acceptance region, are consistent with that fixed value. The confidence set fixes the sample value and asks what parameter values, i.e. the confidence interval, make this sample value most plausible. In short, in the conditions of the previous theorem:

$$(x_1, \ldots, x_n) \in \Omega_\alpha(\theta_0) \quad \text{if and only if} \quad \theta_0 \in \Theta_\alpha(x_1, \ldots, x_n) \ .$$

A final remark to note that even though we are using the theorem above to construct hypothesis tests from confidence set, this result is very often used for the opposite purpose, i.e. to build confidence sets from hypothesis tests.

In the next subsections, we will translate some of the confidence intervals for the parameters of normal populations discussed in the previous section to the setup of hypothesis testing. We conclude with some short comments about tests for fit of a distribution and independency of samples.

4.4.1 Mean

We distinguish between the following two cases:

1) Let $X_1, \ldots, X_n$ be a random sample of size n taken from a normal population $N(\mu, \sigma^2)$ with unknown mean μ and known variance σ^2.
Let z_α be such that $P(Z > z_\alpha) = \alpha$, where $Z \sim N(0, 1)$.
Statistical test with significance $1 - \alpha$ of $H_0 : \mu = \mu_0$ against H_1:

- if $H_1 : \mu \neq \mu_0$: Reject H_0 if $\overline{X}_n < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or $\overline{X}_n > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
- if $H_1 : \mu > \mu_0$: Reject H_0 if $\overline{X}_n > \mu_0 + z_{\alpha}\frac{\sigma}{\sqrt{n}}$.
- if $H_1 : \mu < \mu_0$: Reject H_0 if $\overline{X}_n < \mu_0 - z_{\alpha}\frac{\sigma}{\sqrt{n}}$.

2) Let $X_1, \ldots, X_n$ be a random sample of size n taken from a normal population $N(\mu, \sigma^2)$ with unknown mean μ and variance σ^2.
Let $t_{\alpha,n-1}$ be such that $P(T > t_{\alpha,n-1}) = \alpha$, where $T \sim t(n-1)$.
Statistical test with significance $1-\alpha$ of $H_0 : \mu = \mu_0$ against H_1:

- if $H_1 : \mu \neq \mu_0$: Reject H_0 if $\overline{X}_n < \mu_0 - t_{\alpha/2,n-1}\frac{S'}{\sqrt{n}}$ or $\overline{X}_n > \mu_0 + t_{\alpha/2,n-1}\frac{S'}{\sqrt{n}}$.
- if $H_1 : \mu > \mu_0$: Reject H_0 if $\overline{X}_n > \mu_0 + t_{\alpha,n-1}\frac{S'}{\sqrt{n}}$.
- if $H_1 : \mu < \mu_0$: Reject H_0 if $\overline{X}_n < \mu_0 - t_{\alpha,n-1}\frac{S'}{\sqrt{n}}$.

4.4.2 Variance

Let $X_1, \ldots, X_n$ be a random sample of size n taken from a normal population $N(\mu, \sigma^2)$ with unknown mean μ and variance σ^2.
Let $q_{\alpha,n-1}$ be such that $P(Q > q_{\alpha,n-1}) = \alpha$, where $Q \sim \chi^2(n-1)$.
Statistical test with significance $1-\alpha$ of $H_0 : \sigma^2 = \sigma_0^2$ against H_1:

- if $H_1 : \sigma^2 \neq \sigma_0^2$: Reject H_0 if

$$S'^2 < \frac{q_{1-\alpha/2,n-1}\sigma_0^2}{n-1} \quad \text{or} \quad S'^2 > \frac{q_{\alpha/2,n-1}\sigma_0^2}{n-1} \ .$$

- if $H_1 : \sigma^2 > \sigma_0^2$: Reject H_0 if

$$S'^2 > \frac{q_{\alpha,n-1}\sigma_0^2}{n-1} \ .$$

- if $H_1 : \sigma^2 < \sigma_0^2$: Reject H_0 if

$$S'^2 < \frac{q_{1-\alpha,n-1}\sigma_0^2}{n-1} \ .$$

4.4.3 Difference of two means

We distinguish between the following three cases:

1) Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken, respectively, from two normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with unknown means μ_1 and μ_2 and known variances σ_1^2 and σ_2^2.
Let z_α be such that $P(Z > z_\alpha) = \alpha$, where $Z \sim N(0,1)$.
Statistical test with significance $1-\alpha$ of $H_0 : \mu_1 = \mu_2$ against H_1:

- if $H_1 : \mu_1 \neq \mu_2$: Reject H_0 if

$$\overline{X_{1n}} - \overline{X_{2m}} < -z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \quad or \quad \overline{X_{1n}} - \overline{X_{2m}} > z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \ .$$

- if $H_1 : \mu_1 > \mu_2$: Reject H_0 if

$$\overline{X_{1n}} - \overline{X_{2m}} > z_\alpha \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \ .$$

- if $H_1 : \mu_1 < \mu_2$: Reject H_0 if

$$\overline{X_{1n}} - \overline{X_{2m}} < -z_\alpha \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \ .$$

2) Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken, respectively, from two normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with unknown means μ_1 and μ_2 and unknown, but equal, variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Let $t_{\alpha,n+m-2}$ be such that $P(T > t_{\alpha,n+m-2}) = \alpha$, where $T \sim t(n+m-2)$, and let S^* be given by

$$S^* = \sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_1'^2 + (m-1)S_2'^2}{n+m-2}} \ .$$

Statistical test with significance $1-\alpha$ of $H_0 : \mu_1 = \mu_2$ against H_1:

- if $H_1 : \mu_1 \neq \mu_2$: Reject H_0 if

$$\overline{X_{1n}} - \overline{X_{2m}} < -t_{\alpha/2,n+m-2} S^* \quad \text{or} \quad \overline{X_{1n}} - \overline{X_{2m}} > t_{\alpha/2,n+m-2} S^* \ .$$

- if $H_1 : \mu_1 > \mu_2$: Reject H_0 if $\overline{X_{1n}} - \overline{X_{2m}} > t_{\alpha,n+m-2} S^*$.
- if $H_1 : \mu_1 < \mu_2$: Reject H_0 if $\overline{X_{1n}} - \overline{X_{2m}} < -t_{\alpha,n+m-2} S^*$.

3) Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken, respectively, from two normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with unknown means μ_1 and μ_2 and unknown and unequal variances σ_1^2 and σ_2^2. Let S^* be given by

$$S^* = \sqrt{\frac{S_1'^2}{n} + \frac{S_2'^2}{m}} \ .$$

and let $t_{\alpha,r}$ be such that $P(T > t_{\alpha,r}) = \alpha$, where $T \sim t(r)$ and r is the integer part of r^*:

$$r^* = \frac{\left(\frac{S_1'^2}{n} + \frac{S_2'^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_1'^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_2'^2}{m}\right)^2} \ .$$

Statistical test with significance $1-\alpha$ of $H_0 : \mu_1 = \mu_2$ against H_1:

- if $H_1 : \mu_1 \neq \mu_2$: Reject H_0 if $\overline{X_{1n}} - \overline{X_{2m}} < -t_{\alpha/2,r} S^*$ or $\overline{X_{1n}} - \overline{X_{2m}} > t_{\alpha/2,r} S^*$.
- if $H_1 : \mu_1 > \mu_2$: Reject H_0 if $\overline{X_{1n}} - \overline{X_{2m}} > t_{\alpha,r} S^*$.
- if $H_1 : \mu_1 < \mu_2$: Reject H_0 if $\overline{X_{1n}} - \overline{X_{2m}} < -t_{\alpha,r} S^*$.

4.4.4 Ratio of two variances

Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken, respectively, from two normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with unknown means and variances.
Let $f_{\alpha,n-1,m-1}$ be such that $P(F > f_{\alpha,n-1,m-1}) = \alpha$, where $F \sim F(n-1, m-1)$.
Statistical test with significance $1-\alpha$ of $H_0 : \sigma_1^2 = \sigma_2^2$ against H_1:

- if $H_1 : \sigma_1^2 \neq \sigma_2^2$: Reject H_0 if

$$\frac{S_1'^2}{S_2'^2} < f_{1-\alpha/2,n-1,m-1} \quad \text{or} \quad \frac{S_1'^2}{S_2'^2} > f_{\alpha/2,n-1,m-1} \ .$$

- if $H_1 : \sigma_1^2 > \sigma_2^2$: Reject H_0 if

$$\frac{S_1'^2}{S_2'^2} > f_{\alpha,n-1,m-1} \ .$$

- if $H_1 : \sigma_1^2 < \sigma_2^2$: Reject H_0 if

$$\frac{S_1'^2}{S_2'^2} < f_{1-\alpha,n-1,m-1} \ .$$

4.4.5 Proportion

Let $X_1, \ldots, X_n$ be a random sample of size n taken from a Bernoulli population $Bi(1, p)$ with unknown parameter (proportion) p. Assume that the sample size n is sufficiently large so that the central limit theorem may be used to provide an asymptotic distribution for $\overline{X}_n$.
Let z_α be such that $P(Z > z_\alpha) = \alpha$, where $Z \sim N(0, 1)$.
Statistical test with significance $1-\alpha$ of $H_0 : p = p_0$ against H_1:

- if $H_1 : p \neq p_0$: Reject H_0 if

$$\overline{X}_n < p_0 - z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} \quad \text{or} \quad \overline{X}_n > p_0 + z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} \ .$$

- if $H_1 : p > p_0$: Reject H_0 if

$$\overline{X}_n > p_0 + z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} \ .$$

- if $H_1 : p < p_0$: Reject H_0 if

$$\overline{X}_n < p_0 - z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} \ .$$

4.4.6 Difference of two proportions

Let $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ be two independent random samples of size n and m taken, respectively, from two Bernoulli populations $Bi(1, p_1)$ and $Bi(1, p_2)$ with unknown proportions p_1 and p_2. Assume that the sample sizes n and m are sufficiently large so that the central limit theorem may be used to provide an

asymptotic distribution for $\overline{X_{1n}} - \overline{X_{2m}}$.
Let S^* be given by

$$S^* = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{m}+\frac{1}{n}\right)},$$

where

$$\hat{p} = \frac{n\overline{X_{1n}} + m\overline{X_{2m}}}{n+m},$$

and let z_α be such that $P(Z > z_\alpha) = \alpha$, where $Z \sim N(0,1)$.
Statistical test with significance $1-\alpha$ of $H_0 : p_1 = p_2$ against H_1:

- if $H_1 : p_1 \neq p_2$: Reject H_0 if $\overline{X_{1n}} - \overline{X_{2m}} < -z_{\alpha/2}S^*$ or $\overline{X_{1n}} - \overline{X_{2m}} > z_{\alpha/2}S^*$.
- if $H_1 : p_1 > p_2$: Reject H_0 if $\overline{X_{1n}} - \overline{X_{2m}} > z_\alpha S^*$
- if $H_1 : p_1 < p_2$: Reject H_0 if $\overline{X_{1n}} - \overline{X_{2m}} < -z_\alpha S^*$

4.4.7 Other tests

The tests described above are suitable to test the parameters of normal populations. Moreover, as illustrated with the tests for the proportions of a Bernoulli population, if the random samples is sufficiently large, it is possible to use the central limit theorem to produce tests for the population mean of non-normal populations.

However, very seldom is the probability distribution of the population under observation exactly known. In such cases, the best one can aim for is to make an "informed" guess about that probability distribution and to validate or reject such guess using hypothesis tests. Tests with this goal usually have a null hypothesis stating that the population has a given distribution, while the alternative hypothesis asserts that this is not the case. Examples for this class of tests include chi-square test of goodness of fit and the non-parametric Kolmogorov-Smirnov. Further details can be found in the references.

A key assumption used often in probability and statistics is independence. Given two samples, there may be interest in finding whether the two samples are independent or not, before proceeding further with their statistical treatment. This goal can be achieved through the use of the chi-square test for independency.

Indeed, it should be remarked that there is a very large range of statistical tests, each one adapted to a particular case or situation. Check the references for more information in this topic.

References

[1] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, 2nd edition, 2001.

[2] D.R. Cox and V. Isham. *Point Processes*. Chapman and Hall/CRC, 1980.

[3] J.L. Devore and K.N. Berk. *Modern Mathematical Statistics with Applications*. Springer, 2nd edition, 2012.

[4] G. Grimmett and D. Stirzacker. *Statistical Inference*. Oxford University Press, 2nd edition, 2001.

[5] B. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6th edition, 2003.

[6] P. E. Pfeiffer. *Concepts of Probability Theory*. Dover Publications, 2nd revised edition, 2012.

[7] J. Pitman. *Probability*. Springer texts in Statistics. Springer, 1993.

[8] P. Protter and J. Jacod. *Probability Essentials*. Springer, 2nd edition, 2004.

[9] S. Ross. *Stochastic Processes*. Wiley, 2nd edition, 1995.

[10] S. Ross. *A first course in probability*. Pearson, 8th edition, 2010.

[11] D. Stirzaker. *Probability and Random Variables: A Beginner's Guide*. Cambridge University Press, 1999.

[12] Y. Suhov and M. Kelbert. *Probability and Statistics by Example: Volume 1, Basic Probability and Statistics*. Cambridge University Press, 2005.