A BASIC INTRODUCTION TO PROBABILITY AND STATISTICS FOR MATHEMATICIANS

DANIEL V. TAUSK

Contents

Date: January 24th, 2023.

1. Introduction

Introductory statistics textbooks are usually written for students with little prior knowledge of university-level mathematics, typically just a basic understanding of calculus and matrix algebra. Additionally, abstract linear algebra is often avoided even in more advanced textbooks, with the use of matrices preferred over more abstract concepts such as vector spaces, linear transformations, dual spaces and tensor products. In probability theory, only the so called advanced books make use of the language of abstract measure theory.

Professional mathematicians and graduate mathematics students typically have a strong background in abstract linear algebra and also a reasonable amount of background in abstract measure theory, topology and functional analysis. Although other audiences, who are more interested in applications and have less mathematical background, will understandably try to avoid such abstractions when learning probability and statistics, for mathematicians it is the opposite: these abstractions are familiar topics that are part of their everyday work, and there is no reason to avoid them. In fact, a significant portion (though not all) of the material presented in advanced probability textbooks is already well-known to mathematicians, but with different terminology and motivation. Thus, just by being presented with a translation from measure theory or functional analysis language to probability theory language they will be able to learn a lot about the subject.

This is then the goal of this text: presenting some of the main ideas of probability and statistics to an audience that knows nothing about those subjects but for which the abstract mathematics is the easy part. Also, we will make an effort to explain the main ideas and motivations behind each topic instead of just following the dry definition-lemma-theorem-proof style used by typical mathematics books.

2. Probability spaces and events

A probability space is simply a measure space in which the measure of the entire space is equal to 1. We briefly recall the relevant measure-theoretic concepts just to make sure our terminology is clear.

Definition 2.1. A *probability space* is a triple $(\Omega, \mathcal{A}, \mathbb{P})$, where Ω is a set, $\mathcal{A}$ is a σ-algebra of subsets of Ω (i.e., a nonempty collection of subsets of Ω closed under countable unions and complements) and $\mathbb{P}$ is a *probability measure* on $\mathcal{A}$, i.e., a countably additive nonnegative measure defined on $\mathcal{A}$ with $\mathbb{P}(\Omega) = 1$. Elements of $\mathcal{A}$ will be, as usual, called *measurable subsets* of Ω but in the context of probability theory they are also called *events*; for $A \in \mathcal{A}$, we call $\mathbb{P}(A)$ the *probability* of the event A.

Let us discuss a bit the connection between the definition above and practical applications. A probability space can be used as a mathematical model for the set of possible outcomes of a random experiment, what is usually called the *sample space* for that experiment. By a *random experiment* it is meant some procedure that can be repeated as often as one wishes, with the repetitions being independent of each other. This is not supposed to be a formal mathematical definition in any way, of course, as we are now talking about something outside of the domain of pure mathematics.

One paradigmatic textbook example of a random experiment is a toss of a coin or a throw of the dice. The outcome of the experiment would be either heads or tails in the case of a coin and an integer number between 1 and 6 in the case of a die. More relevant examples, connected to real-world applications of statistics, would be for instance gathering a sample of people from the population and collecting answers from them using a questionnaire or gathering a sample of sick patients and testing a new treatment, collecting data such as the evolution of symptom severity.

If $(\Omega, \mathcal{A}, \mathbb{P})$ is the probability space used to model a given random experiment then each time the experiment is performed an outcome $\omega \in \Omega$ is obtained. An element $A \in \mathcal{A}$ is called an "event" because we are thinking that "the event A happened" is a short for "the event that the obtained outcome ω belongs to A happened". Thus, $\mathbb{P}(A)$ — the probability of the event A — is understood as the probability that the obtained outcome ω is in A. A possible interpretation for the number $\mathbb{P}(A)$, the so called *frequentist* interpretation, is that $\mathbb{P}(A)$ is the frequency of occurrences of $\omega \in A$ when the random experiment is repeated a large number N of times. In more mathematical language, this can be expressed as the limit of the quotients $\frac{N_A}{N}$ as N tends to $+\infty$, where N_A denotes the number of times among the N repetitions in which the event $\omega \in A$ occurred.

Almost everything we said above after Definition 2.1 is problematic and cannot be taken too seriously. To begin with, it is not completely clear what one means by "repeating" an experiment as, for instance, when we toss a coin a second time we usually don't even make an effort to position our hand in the exact same way as in the first toss. So "repeating" is not

supposed to mean that everything is exactly the same. The possibility of repeating the experiment as often as one wishes is also not supposed to be interpreted *stricto sensu*, as in practice there is obviously some finite (though maybe very large) limit to the number of possible repetitions. The "independence" assumption means that whatever happened in the previous instances of the experiment should not influence the outcome of the current one, though that might not be exactly true in all cases. One should also notice that statistics is often used to analyse observational data, i.e., data that is collected retrospectively and it is not the outcome of some deliberate planned or controlled experiment. It is even less clear what "repeating" would mean in this context.

The frequentist interpretation of probability is also not the only one, as for instance there is also *Bayesian statistics*. In the Bayesian framework, probability is used to express uncertainty about facts due to incomplete knowledge. For instance, one might be willing to talk about the probability that the 100-th digit of π be greater than 4. There is definitely no conceivable sense in which this can be seen as related to repetitions of some procedure, as the 100-th digit of π is simply some fixed number. Nevertheless, if you find yourself in a casino and someone proposes to you a bet based on the value of the 100-th digit of π and if you don't have access to a computer or any means to find out the correct value, you would likely appeal to some sort of probabilistic reasoning to decide if it is a good idea to accept the bet — say, by assigning a probability of $\frac{1}{10}$ to each of the 10 possibilities.

A word of caution should be said with respect to the use of the word "random". Is there anything like true randomness in nature and what does that mean exactly? Is the outcome of a certain experiment really random or maybe it was determined from complete specification of initial conditions and the laws of physics? None of this matters for statistics. The fact is that often the outcomes of sufficiently complicated deterministic processes exhibit statistical regularities, i.e., properties that one can successfully study using the methods of probability theory. For example, a pseudo-random number generator is certainly a deterministic process and yet one can make good predictions about the frequency properties of its long term outcomes using probability theory. So statistical reasoning can be thought of as an approach to studying complicated phenomena — whether true randomness is involved or not — in which complicated details (say, microscopic details, uncontrollable variables, etc) are ignored.

Let us now go back to the mathematics to make a few observations and look at a couple of simple examples. We note that if $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space, $A \subset \Omega$ is a countable set and every singleton $\{\omega\}$ with $\omega \in A$ belongs to the σ-algebra $\mathcal{A}$, then $A \in \mathcal{A}$ and the probability of A is given by

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega), \tag{2.1}$$

where we abbreviate $\mathbb{P}\big(\{\omega\}\big)$ as $\mathbb{P}(\omega)$. Therefore, if Ω itself is countable and all the singletons $\{\omega\}$, $\omega \in \Omega$, belong to $\mathcal{A}$ — which is typically the case — then $\mathcal{A}$ coincides with the collection $\wp(\Omega)$ of all subsets of Ω and formula (2.1) holds for every subset A of Ω. In this case, the entire probability measure $\mathbb{P}$ is completely determined by the probabilities $\mathbb{P}(\omega)$ of the individual points $\omega \in \Omega$. In fact, any specification of nonnegative real numbers $\mathbb{P}(\omega)$, for all $\omega \in \Omega$, with $\sum_{\omega\in\Omega} \mathbb{P}(\omega) = 1$ yields a unique probability measure $\mathbb{P}$ on the σ-algebra $\mathcal{A} = \wp(\Omega)$ of all subsets of Ω defined by (2.1). A probability space whose underlying set Ω is countable is usually said to be *discrete*.

If Ω is uncountable, formula (2.1) usually does not hold for all $A \in \mathcal{A}$ and in fact it often happens that $\mathbb{P}(\omega) = 0$ for all $\omega \in \Omega$. For example, one can take $\Omega = [0, 1]$ and $\mathbb{P}$ the Lebesgue measure defined on the σ-algebra $\mathcal{A}$ of Lebesgue measurable subsets of $[0, 1]$. As it is well known, for uncountable Ω there are obstructions to defining interesting measures on the collection of all subsets of Ω and that is the main reason why one needs σ-algebras in measure theory. In the discrete case, as discussed above, one often simply takes $\mathcal{A} = \wp(\Omega)$ (though see Subsection 19.1 for another reason why nontrivial σ-algebras are useful even in the discrete case).

We conclude the section with one comment concerning events of probability zero. One would normally expect that an event having zero probability is an impossible event, but as we saw above there are situations in which Ω is uncountable and every $\omega \in \Omega$ has zero probability. Of course, it cannot be true that every $\omega \in \Omega$ is impossible, as Ω is the set of all possible outcomes so that some $\omega \in \Omega$ will be the outcome. Events with zero probability are usually referred to as "almost impossible", with the only truly impossible event being the empty set. In fact, in the context of probability theory the expression *almost surely* or $\mathbb{P}$-*almost surely* is used with the same meaning that the expression "almost everywhere" is used in measure theory. Namely, something happens almost surely if the probability that it doesn't happen is zero or, equivalently, if the probability that it does happen is equal to one.

Though an event with positive probability being a union of (uncountably many) events with zero probability is not a problem within the mathematical formalism, this possibility seems paradoxical when connections to the real world are considered. Note, however, that the set of outcomes of a real experiment — the kind of outcome that you would obtain using some kind of measuring apparatus and then write down in a piece of paper or in a computer spreadsheet — is always finite. Laboratory equipment and computers have limited precision and the set of all real numbers in some interval is never going to be the set of all possible outcomes. The reason why we use nondiscrete probability spaces is not that they model the outcomes of experiments more faithfully, it is because they make the mathematics simpler and more elegant.

3. Random variables and random objects

Though a point of the underlying set Ω of a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is denoted by a simple innocent letter like ω, in many concrete applications such a point will encode *a lot* of information. For example, one can consider an experiment in which a large sample of the population of a country is sampled and a lot of data (for instance, age, height, weight, home address, health data, etc) from the individuals of that sample is collected. One would then consider a probability space such that each $\omega \in \Omega$ contains all the collected data from all the individuals in that sample. Or maybe a point ω could represent all the data from the stock market during a certain period of time. Mathematically speaking, ω would typically be represented as an element of some large cartesian product of sets. Out of the large dataset that a single point ω of Ω represents, one will usually want to isolate specific quantities of interest $X(\omega)$ which will be discussed and used in computations. Such quantities are often real-valued.

For example, if ω contains all the data collected from a sample of the population of a country then $X(\omega)$ could be the average of the heights of the individuals in that sample, or the median of the ages of such individuals, or the weight of the seventh individual in the sample (according to some specified ordering) or the number of individuals in the sample that live in a certain region of the country. If ω contains all the data from the stock market during a certain period of time then $X(\omega)$ could be the value of some particular stock in some particular moment — and so on.

In mathematical terms, a real-valued quantity of interest associated to points of Ω would be represented by a map $X : \Omega \to \mathbb{R}$. One would then be interested in asking questions like "what is the probability that the value of X belongs to a given subset B of $\mathbb{R}$?" or, for a more specific example, "what is the probability that the value of X be greater than 1?". The probability that the value of X belongs to B, denoted by $\mathbb{P}(X \in B)$, should of course be defined as the probability of the set

$$X^{-1}[B] = \big\{\omega \in \Omega : X(\omega) \in B\big\}$$

which is just the inverse image of B under the map X. In other words, we define:

$$\mathbb{P}(X \in B) = \mathbb{P}\big(X^{-1}[B]\big). \tag{3.1}$$

It is also convenient to denote the set $X^{-1}[B]$ by $[X \in B]$ and, in the same spirit, to use notations such as $[X > x]$ and $\mathbb{P}(X > x)$ with the obvious meaning.

In order for the probability (3.1) to make sense it is necessary that $X^{-1}[B]$ belongs to the σ-algebra $\mathcal{A}$. In most cases it is not reasonable to expect this to happen for arbitrary subsets B of $\mathbb{R}$ but one would hope this to happen at least for sufficiently simple subsets of $\mathbb{R}$, say, for intervals. As intervals generate the entire σ-algebra of Borel subsets of $\mathbb{R}$, if $X^{-1}[B] \in \mathcal{A}$ whenever B is an interval, it will also be the case that $X^{-1}[B] \in \mathcal{A}$ for every Borel

subset B of $\mathbb{R}$. What we are saying here is that the map $X : \Omega \to \mathbb{R}$ should be measurable.

We quickly recall the relevant measure-theoretic definitions.

Definition 3.1. A *measurable space* $(M, \mathcal{B})$ is a set M endowed with a σ-algebra $\mathcal{B}$ of subsets of M. The elements of $\mathcal{B}$ are called *measurable subsets* of M. Given measurable spaces $(M, \mathcal{B})$ and $(M', \mathcal{B}')$, a map $f : M \to M'$ is said to be *measurable* if $f^{-1}[B] \in \mathcal{B}$ for all $B \in \mathcal{B}'$.

Clearly, for $f : M \to M'$ to be measurable, it is sufficient that $f^{-1}[B] \in \mathcal{B}$ for all B in a collection of generators for the σ-algebra $\mathcal{B}'$.

We now give the main definition of the section.

Definition 3.2. A *random variable* X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a measurable function $X : \Omega \to \mathbb{R}$, where Ω is endowed with the σ-algebra $\mathcal{A}$ and $\mathbb{R}$ is endowed with its Borel σ-algebra.

Example 3.3. Here is a dumb example so that we can introduce the relevant terminology. Given a subset A of a set Ω, the function $\mathbf{1}_A : \Omega \to \mathbb{R}$ that takes the value 1 on A and the value zero on $\Omega \setminus A$ is called the *indicator function* of A. Mathematicians usually call this the characteristic function of A, but in probability theory the name "characteristic function" is reserved for something else (see Section 14), so one uses "indicator function" instead. If $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space and $A \in \mathcal{A}$ is an event then obviously $\mathbf{1}_A$ is a random variable in $(\Omega, \mathcal{A}, \mathbb{P})$.

Though many quantities of interest are real-valued, this is not always the case and nothing stops us from considering the following obvious generalization of the concept of random variable.

Definition 3.4. Given a measurable space $(M, \mathcal{B})$, by an $(M, \mathcal{B})$-*valued random object* X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ we mean a measurable map $X : \Omega \to M$. For every $B \in \mathcal{B}$, we write $\mathbb{P}(X \in B) = \mathbb{P}\big(X^{-1}[B]\big)$. If $M = \mathbb{R}^n$ (or if M is a real finite dimensional vector space) and $\mathcal{B}$ is the Borel σ-algebra of M then an $(M, \mathcal{B})$-valued random object is also called a *random vector*.

Recall that for an arbitrary topological space its *Borel σ-algebra* is defined as the σ-algebra generated by the open sets and that every real finite-dimensional vector space has a canonical topology (which can be defined, for instance, as the topology induced by an arbitrary norm). Clearly, an $\mathbb{R}^n$-valued random vector X is the same as an n-tuple $(X_1, \ldots, X_n)$ of random variables, as the measurability of an $\mathbb{R}^n$-valued map is equivalent to its coordinatewise measurability.

The language of random variables is very convenient as it matches closely the way statisticians think when handling concrete problems. For example, one can use various operations with random variables to create new random variables. Say, if X, Y, Z and W are random variables on the same probability space and W never vanishes, one can construct a new random

variable by using a formula like $\frac{1}{W}(X^2Y - 3Z)$. Such operations with random variables are to be understood as one usually understands operations with functions having the same domain, i.e., operations are defined pointwise. More generally, one can apply a function f to a random variable X forming a new random variable $f(X)$. What one is thinking here is that f is applied to some observed value $X(\omega)$ of the random variable X, so $f(X)$ should be understood as a composition.

Definition 3.5. If $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ are measurable spaces, $f : M \to M'$ is a measurable map and X is an $(M, \mathcal{B})$-valued random object on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, then we define $f(X)$ as the $(M', \mathcal{B}')$-valued random object on that same probability space given by $f(X) = f \circ X$.

Random variables are "variables" not in the sense that logic textbooks use the term, but in the sense that physicists (and pre-twentieth century mathematicians) use the term: they are quantities of interest that might be related with each other by functions. The "variables" used by physicists could be mathematically formalized as (usually real-valued) functions defined in some state space and functions of such variables would be defined, as above, using compositions (see [7] for a more complete discussion). In statistics, we simply replace the state space with a sample space carrying a probability measure and we imagine that the point of the sample space was obtained through some random process. In this sense the point ω becomes random and thus also the value $X(\omega)$ of X becomes random, i.e., the variable X inherits the randomness from its domain, hence random variable.

4. The distribution of a random object

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to M$ be a random object, where $(M, \mathcal{B})$ is a measurable space. The *probability distribution* (or simply *distribution*) of the random object X is the probability measure $\mathbb{P}_X$ on the σ-algebra $\mathcal{B}$ defined by

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}\big(X^{-1}[B]\big),$$

for all $B \in \mathcal{B}$.

The construction above is actually familiar from abstract measure theory and it is known as the push-forward (or image) of a measure under a map. Namely, if f is a measurable function between measurable spaces and μ is a measure defined on the σ-algebra in the domain of f then the *push-forward* of μ under f is the measure $f_*\mu$ defined on the σ-algebra in the counter-domain of f given by $(f_*\mu)(B) = \mu\big(f^{-1}[B]\big)$, for every B in that σ-algebra. Hence the distribution of a random object X is simply the push-forward under X of the probability measure in the domain of X:

$$\mathbb{P}_X = X_* \, \mathbb{P}.$$

Clearly, if two random objects $X : \Omega \to M$, $Y : \Omega \to M$ are equal almost surely then $\mathbb{P}_X = \mathbb{P}_Y$.

Example 4.1. A random object X is called *discrete* if its image $\mathrm{Im}(X)$ is a countable set. If X is a discrete $(M, \mathcal{B})$-valued random object such that the singleton $\{x\}$ is in $\mathcal{B}$ for every x in the image of X then the distribution of X is completely determined by the values of $\mathbb{P}_X\big(\{x\}\big)$ — abbreviated as $\mathbb{P}_X(x)$ — with x in the image of X. Namely, we have

$$\mathbb{P}_X(B) = \sum_{x \in B \cap \mathrm{Im}(X)} \mathbb{P}_X(x),$$

for any $B \in \mathcal{B}$.

Note that if $X : \Omega \to M$ is a random object and $f : M \to M'$ is a measurable map taking values in some other measurable space $(M', \mathcal{B}')$ then the distribution of the random object $f(X)$ is simply the push-forward under f of the distribution of X:

$$\mathbb{P}_{f(X)} = (f \circ X)_* \mathbb{P} = f_* X_* \mathbb{P} = f_* \, \mathbb{P}_X.$$

In particular, the distribution of $f(X)$ depends only on the distribution of X. This is a useful observation, as several probability distributions that are important in statistics are defined by a statement of the form "it is the distribution of $f(X)$ for a certain given f, where the distribution of X is ...". The previous observation implies that this type of definition is valid, as the distribution of $f(X)$ does not depend on the choice of the random object X as long as X has some required distribution.

5. A useful technical lemma

In order to check that two probability measures defined in the same σ-algebra are equal it is not sufficient to check that they agree in a collection of generators of the σ-algebra. For example, if $\Omega = \{0, 1, 2, 3\}$ then $\big\{\{0, 1\}, \{1, 2\}\big\}$ generates $\wp(\Omega)$ and yet it is easy to give examples of distinct probability measures on $\wp(\Omega)$ that have the same value on the sets $\{0, 1\}$ and $\{1, 2\}$. We can fix this by adding a simple hypothesis to the set of generators.

Lemma 5.1. *Let $(\Omega, \mathcal{A})$ be a measurable space and let μ and ν be finite countably additive measures defined on $\mathcal{A}$. Let $\mathcal{C} \subset \mathcal{A}$ be a collection that generates the σ-algebra $\mathcal{A}$ and is closed under finite intersections. If*

$$\mu(A) = \nu(A)$$

for all $A \in \mathcal{C} \cup \{\Omega\}$ then $\mu = \nu$. In particular, if two probability measures defined on $\mathcal{A}$ coincide on elements of $\mathcal{C}$ then they are equal.

We will prove Lemma 5.1 in a moment, but first we need a definition and another lemma. The reason why the obvious approach for proving Lemma 5.1 fails is that the collection of sets in which two probability measures coincide is not a σ-algebra. Nevertheless, the collection of sets in which two finite countably additive measures coincide does satisfy certain closure properties and this leads us to the notion of σ-additive class. A nonempty

collection $\mathcal{S}$ of sets is called a *σ-additive class* if it is closed under finite disjoint unions (i.e., $A, B \in \mathcal{S}$ and $A \cap B = \emptyset$ implies $A \cup B \in \mathcal{S}$), proper differences (i.e., $A, B \in \mathcal{S}$ and $B \subset A$ implies $A \setminus B \in \mathcal{S}$) and increasing limits (i.e., if $(A_n)_{n\geq 1}$ is an increasing sequence of sets in $\mathcal{S}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{S}$).

We recall that a *σ-ring* is a nonempty collection of sets that is closed under countable unions and differences. Note that a σ-algebra of subsets of Ω is the same as a σ-ring consisting of subsets of Ω of which Ω itself is a member. Moreover, one easily checks that a σ-additive class closed under finite intersections is a σ-ring.

Lemma 5.2. *If $\mathcal{C}$ is a collection of sets closed under finite intersections then the σ-additive class generated by $\mathcal{C}$ coincides with the σ-ring generated by $\mathcal{C}$.*

Proof. It is sufficient to check that the σ-additive class $\mathcal{S}$ generated by $\mathcal{C}$ is closed under finite intersections. To this aim, check first that for any set A, the collection

$$\big\{B \in \mathcal{S} : A \cap B \in \mathcal{S}\big\} \tag{5.1}$$

is a σ-additive class. For $A \in \mathcal{C}$, the collection (5.1) contains $\mathcal{C}$ and thus it contains $\mathcal{S}$. This establishes that the intersection of a member of $\mathcal{C}$ with a member of $\mathcal{S}$ is in $\mathcal{S}$. Now repeat this reasoning noting that we have just proven that the collection (5.1) contains $\mathcal{C}$ for any $A \in \mathcal{S}$. □

Proof of Lemma 5.1. Use Lemma 5.2 keeping in mind that

$$\big\{A \in \mathcal{A} : \mu(A) = \nu(A)\big\}$$

is a σ-additive class that contains the collection $\mathcal{C} \cup \{\Omega\}$ which is closed under finite intersection and generates $\mathcal{A}$ as a σ-ring. □

6. Joint distributions and marginals

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Given random objects X and Y on $(\Omega, \mathcal{A}, \mathbb{P})$ taking values in measurable spaces $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ respectively, we can form a new random object $(X, Y) : \Omega \to M \times M'$ by setting

$$(X, Y)(\omega) = \big(X(\omega), Y(\omega)\big),$$

for all $\omega \in \Omega$. The set $M \times M'$ should be endowed with the *product σ-algebra* $\mathcal{B} \otimes \mathcal{B}'$ which is the σ-algebra generated by all products $B \times B'$ with $B \in \mathcal{B}$ and $B' \in \mathcal{B}'$. Such σ-algebra is appropriate since it has the property that (X, Y) is measurable if and only if both X and Y are measurable.

Definition 6.1. The distribution $\mathbb{P}_{(X,Y)}$ of the random object (X, Y) is known as the *joint distribution* of X and Y.

Note that if

$$\pi_1 : M \times M' \to M, \quad \pi_2 : M \times M' \to M'$$

denote the projections then

$$\mathbb{P}_X = (\pi_1)_* \mathbb{P}_{(X,Y)}, \quad \mathbb{P}_Y = (\pi_2)_* \mathbb{P}_{(X,Y)},$$

i.e., the distributions of X and Y are obtained by taking the push-forward of the joint distribution under the projections. In this context one usually says that the distributions of X and Y are the *marginal* distributions corresponding to the joint distribution $\mathbb{P}_{(X,Y)}$. Note, however, that this is just a new name for the distributions of X and Y.

To motivate such terminology let us look at the case in which the sets M and M' are countable and $\mathcal{B} = \wp(M)$, $\mathcal{B}' = \wp(M')$, so that $M \times M'$ is also countable and $\mathcal{B} \otimes \mathcal{B}' = \wp(M \times M')$. As discussed in Section 2, the probability measure in a discrete (i.e., countable) probability space is determined by the probabilities of individual elements and thus if X is an $(M, \mathcal{B})$-valued and Y is an $(M', \mathcal{B}')$-valued random object then the joint distribution of X and Y is determined by probabilities of the form:

$$\mathbb{P}_{(X,Y)}(x, y) = \mathbb{P}\big([X = x] \cap [Y = y]\big),$$

with $x \in M$ and $y \in M'$. The distributions of X and Y are then obtained by taking the sums

$$\mathbb{P}_X(x) = \sum_{y \in M'} \mathbb{P}_{(X,Y)}(x, y), \quad \mathbb{P}_Y(y) = \sum_{x \in M} \mathbb{P}_{(X,Y)}(x, y), \tag{6.1}$$

for all $x \in M$ and all $y \in M'$. We normally imagine the probabilities $\mathbb{P}_{(X,Y)}(x, y)$ written in a rectangular table and the row and column totals $\mathbb{P}_X(x)$ and $\mathbb{P}_Y(y)$ written on the margins of such table — hence *marginal* distributions.

The notion of joint distribution can be generalized to arbitrary families of random objects. We recall some definitions.

Definition 6.2. Given a set M, a family $\big((M_i, \mathcal{B}_i)\big)_{i \in I}$ of measurable spaces and a family $(f_i)_{i \in I}$ of maps $f_i : M \to M_i$, we define the σ-algebra of subsets of M *induced* by the family $(f_i)_{i \in I}$ as the smallest σ-algebra which makes all of the maps f_i measurable. This obviously coincides with the σ-algebra generated by:

$$\bigcup_{i \in I} \big\{ f_i^{-1}[B] : B \in \mathcal{B}_i \big\}.$$

Note that if $\mathcal{C}_i$ generates the σ-algebra $\mathcal{B}_i$ for each $i \in I$ then

$$\bigcup_{i \in I} \big\{ f_i^{-1}[B] : B \in \mathcal{C}_i \big\} \tag{6.2}$$

generates the σ-algebra induced by the family $(f_i)_{i \in I}$; namely, note that all maps f_i are measurable with respect to the σ-algebra generated by (6.2).

It is readily checked that if M is endowed with the σ-algebra induced by $(f_i)_{i \in I}$ and g is an M-valued map defined in some arbitrary measurable space then g is measurable if and only if $f_i \circ g$ is measurable, for all $i \in I$.

Definition 6.3. Let $\big((M_i, \mathcal{B}_i)\big)_{i\in I}$ be a family of measurable spaces. The σ-algebra of subsets of the cartesian product $M = \prod_{i\in I} M_i$ induced by the projections

$$\pi_i : M \longrightarrow M_i, \quad i \in I$$

is denoted by $\bigotimes_{i\in I} \mathcal{B}_i$ and it is called the *product* σ-algebra.

Clearly, an M-valued map g is measurable if and only if all of its coordinates $\pi_i \circ g$ are measurable.

Definition 6.4. If $(X_i)_{i\in I}$ is a family of random objects on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with X_i taking values in a measurable space $(M_i, \mathcal{B}_i)$ then the *joint* distribution of the family $(X_i)_{i\in I}$ is the distribution $\mathbb{P}_{(X_i)_{i\in I}}$ of the $\big(\prod_{i\in I} M_i, \bigotimes_{i\in I} \mathcal{B}_i\big)$-valued random object defined by

$$(X_i)_{i\in I}(\omega) = \big(X_i(\omega)\big)_{i\in I} \in \prod_{i\in I} M_i, \tag{6.3}$$

for all $\omega \in \Omega$.

We will use the same notation for the family of maps $(X_i)_{i\in I}$ and for the map (6.3). This should not cause much confusion.

As in the case of two random objects, the individual distributions of the random objects X_i can be obtained from the joint distribution of the family by taking push-forwards under the projections and in this context we call such individual distributions *marginal* distributions.

Note that if, for each $i \in I$, $\mathcal{C}_i$ is a collection of generators for $\mathcal{B}_i$ that is closed under finite intersections than the collection of all sets of the form

$$\pi_{i_1}^{-1}[C_1] \cap \ldots \cap \pi_{i_n}^{-1}[C_n], \quad C_1 \in \mathcal{C}_{i_1},\ \ldots,\ C_n \in \mathcal{C}_{i_n}, \tag{6.4}$$

with $i_1, \ldots, i_n \in I$ distinct and $n \geq 1$ generates $\bigotimes_{i\in I} \mathcal{B}_i$ and it is closed under finite intersections. An application of Lemma 5.1 then yields the following result.

Proposition 6.5. *Let $\big((M_i, \mathcal{B}_i)\big)_{i\in I}$ be a family of measurable spaces and for each $i \in I$ let $\mathcal{C}_i$ be a collection of generators for the σ-algebra $\mathcal{B}_i$ that is closed under finite intersections. We have that two probability measures on the product σ-algebra $\bigotimes_{i\in I} \mathcal{B}_i$ are equal if they coincide on sets of the form* (6.4) *for any $i_1, \ldots, i_n \in I$ distinct and any $n \geq 1$.* □

Proposition 6.5 says that the joint distribution of a family $(X_i)_{i\in I}$ of random objects is completely determined by probabilities of the form

$$\mathbb{P}\big([X_{i_1} \in C_1] \cap \ldots \cap [X_{i_n} \in C_n]\big), \quad C_1 \in \mathcal{C}_{i_1},\ \ldots,\ C_n \in \mathcal{C}_{i_n},$$

with $i_1, \ldots, i_n \in I$ distinct and $n \geq 1$, assuming that X_i is $(M_i, \mathcal{B}_i)$-valued and that $\mathcal{C}_i$ is a collection of generators for $\mathcal{B}_i$ that is closed under finite intersections for all $i \in I$. Taking $\mathcal{C}_i = \mathcal{B}_i$ we conclude in particular that the joint distribution of $(X_i)_{i\in I}$ is completely determined by all the joint distributions of the finite subfamilies $(X_i)_{i\in F}$, with F ranging over the finite subsets of I.

For finite families, Proposition 6.5 can be restated in the following more convenient form.

Proposition 6.6. *Let $\big((M_i, \mathcal{B}_i)\big)_{i\in I}$ be a finite family of measurable spaces and for each $i \in I$ let $\mathcal{C}_i$ be a collection of generators for the σ-algebra $\mathcal{B}_i$ that is closed under finite intersections. We have that two probability measures on the product σ-algebra $\bigotimes_{i\in I} \mathcal{B}_i$ are equal if they coincide on sets of the form*

$$\prod_{i\in I} C_i, \tag{6.5}$$

with $C_i \in \mathcal{C}_i \cup \{M_i\}$ for all $i \in I$. □

Note that in Proposition 6.6 it is crucial to allow the possibility that $C_i = M_i$, otherwise the product sets (6.5) might not generate the product σ-algebra. In Proposition 6.5 it wasn't necessary to allow explicitly for the possibility that $C_i = M_i$ because the collection of indices $\{i_1, \dots, i_n\}$ appearing in (6.4) is allowed to be a proper subset of I even if I is finite.

In probability theory, most relevant questions concerning a family of random objects $(X_i)_{i\in I}$ depend only on their joint probability distribution — in some cases the image of the map (6.3) is also important. In any case, the common probability space $(\Omega, \mathcal{A}, \mathbb{P})$ in which all of the random objects of interest are defined is not important and that is why authors seldom care to clearly specify it. However, one does need to worry about existence results, i.e., results that ensure the existence of a probability space in which a family $(X_i)_{i\in I}$ of random objects with the desired joint distribution (sometimes also with constraints on the image of (6.3)) can be defined. Examples of such existence results will be discussed in Sections 16 and 18.

Remark 6.7. A word of caution must be said about products of σ-algebras. Typically, the spaces M_i above will be topological spaces and $\mathcal{B}_i$ will be the corresponding Borel σ-algebras. The product $M = \prod_{i\in I} M_i$ thus has a product topology and one might be tempted to confuse the product σ-algebra $\mathcal{B} = \bigotimes_{i\in I} \mathcal{B}_i$ with the Borel σ-algebra of the product topology. While the Borel σ-algebra of the product topology always contains the product σ-algebra, the two might differ even in the simplest case of a product of two spaces. When they do not coincide, one cannot infer, for example, the measurability of a function $f\big((X_i)_{i\in I}\big)$ of the random objects X_i from the measurability of the X_i and the continuity of f, as the continuous function f defined on $\prod_{i\in I} M_i$ is measurable with respect to the Borel σ-algebra and the $\prod_{i\in I} M_i$-valued map $(X_i)_{i\in I}$ is measurable with respect to the product σ-algebra. However, if I is countable and the topology of each M_i is second countable then it is easily checked that the Borel σ-algebra of the product does coincide with the product of the Borel σ-algebras of the spaces M_i. Namely, in this case the product topology is also second countable and therefore every open set is a countable union of basic open sets that belong to the product σ-algebra.

7. Cumulative distribution function of a random variable

The distribution of a random variable X is a Borel probability measure $\mathbb{P}_X$ on the real line, i.e., a probability measure defined on the Borel σ-algebra of the real line. Note that the collection $\big\{\,]-\infty, x] : x \in \mathbb{R}\big\}$ is clearly closed under finite intersections and it generates the Borel σ-algebra of $\mathbb{R}$. Thus, by Lemma 5.1, two Borel probability measures on $\mathbb{R}$ that agree on sets of the form $]-\infty, x]$ must be equal. Let then $\mathbb{P}$ be a Borel probability measure on $\mathbb{R}$ and define $F : \mathbb{R} \to \mathbb{R}$ by setting

$$F(x) = \mathbb{P}\big(\,]-\infty, x]\big), \tag{7.1}$$

for all $x \in \mathbb{R}$. Clearly the map F satisfies the following conditions:

(i) F is increasing;
(ii) F is right-continuous;
(iii) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$.

Does every map $F : \mathbb{R} \to \mathbb{R}$ satisfying (i), (ii) and (iii) arises from a Borel probability measure on $\mathbb{R}$? The answer is affirmative and this follows from the following standard result.

Proposition 7.1. *Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing right-continuous function. There exists a unique nonnegative countably additive Borel measure μ on $\mathbb{R}$ such that*

$$\mu\big(\,]a, b]\big) = F(b) - F(a),$$

for all $a, b \in \mathbb{R}$ with $a \le b$. □

Taking Lemma 5.1 and Proposition 7.1 together we obtain the following characterization of Borel probability measures on the real line.

Proposition 7.2. *The mapping $\mathbb{P} \mapsto F$, with F defined as in* (7.1)*, is a bijection between Borel probability measures on $\mathbb{R}$ and functions $F : \mathbb{R} \to \mathbb{R}$ satisfying (i), (ii) and (iii) above.* □

Proposition 7.2 motivates the following definition.

Definition 7.3. Let X be a random variable. The *cumulative distribution function* (cdf) of X is the map $F_X : \mathbb{R} \to \mathbb{R}$ defined by

$$F_X(x) = \mathbb{P}_X\big(\,]-\infty, x]\big) = \mathbb{P}(X \le x),$$

for all $x \in \mathbb{R}$.

In other words, F_X is the map corresponding to the Borel probability measure $\mathbb{P}_X$ under the bijection given by Proposition 7.2. It follows that a map $F : \mathbb{R} \to \mathbb{R}$ is the cumulative distribution function of some random variable if and only if F satisfies (i), (ii) and (iii) above.

8. Probability density functions

Let $(M, \mathcal{B})$ be a measurable space and let μ and ν be countably additive nonnegative measures defined on $\mathcal{B}$. Recall that ν is said to be *absolutely continuous* with respect to μ if $\mu(B) = 0$ implies $\nu(B) = 0$ for all $B \in \mathcal{B}$. Assume that μ and ν are both σ-finite; recall that a measure defined on $\mathcal{B}$ is called *σ-finite* if M is a countable union of sets of $\mathcal{B}$ having finite measure. The celebrated *Radon–Nikodym Theorem* states that ν is absolutely continuous with respect to μ if and only if there exists a nonnegative real-valued measurable function f defined on M such that

$$\nu(B) = \int_B f \, \mathrm{d}\mu,$$

for all $B \in \mathcal{B}$. The function f is unique up to μ-almost everywhere equality and it is called a *Radon–Nikodym derivative* of ν with respect to μ. A Radon–Nikodym derivative of ν with respect to μ is usually denoted by $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

Definition 8.1. Let X be a random object taking values in a measurable space $(M, \mathcal{B})$ and assume that μ is a σ-finite countably additive nonnegative measure defined on $\mathcal{B}$. If $\mathbb{P}_X$ is absolutely continuous with respect to μ then a Radon–Nikodym derivative of $\mathbb{P}_X$ with respect to μ is called a *probability density function* (pdf) of X with respect to μ. In other words, a probability density function of X with respect to μ is a nonnegative real-valued measurable function $f_X : M \to [0, +\infty[$ such that

$$\mathbb{P}(X \in B) = \mathbb{P}_X(B) = \int_B f_X \, \mathrm{d}\mu,$$

for all $B \in \mathcal{B}$.

Of course, if f_X is the probability density function with respect to μ of some $(M, \mathcal{B})$-valued random object then the integral of f_X with respect to μ must be equal to 1, as $\mathbb{P}_X(M) = 1$. Moreover, every nonnegative measurable map $f_X : M \to [0, +\infty[$ whose integral with respect to μ is equal to 1 yields a probability measure on $\mathcal{B}$ by integration with respect to μ and thus it is the probability density function with respect to μ of some $(M, \mathcal{B})$-valued random object.

Let us look at a few important examples.

Example 8.2. Let X be a random object taking values in a measurable space $(M, \mathcal{B})$. If M is countable, $\mathcal{B} = \wp(M)$ and μ is the counting measure on $\mathcal{B}$ (i.e., $\mu(B)$ is the number of elements of B for all $B \subset M$) then $\mathbb{P}_X$ is always absolutely continuous with respect to μ and the unique probability density function for X with respect to μ is the map $f_X : M \to [0, 1]$ defined by

$$f_X(x) = \mathbb{P}_X(x) = \mathbb{P}(X = x),$$

for all $x \in M$. The map f_X is called the *probability mass function* of X.

Recall that a map $F : [a, b] \to \mathbb{R}$ is called *absolutely continuous* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\sum_{i=1}^{k} |F(y_i) - F(x_i)| < \varepsilon,$$

for any finite collection $]x_1, y_1[, \ldots,]x_k, y_k[$ of disjoint open intervals contained in $[a, b]$ with $\sum_{i=1}^{k}(y_i - x_i) < \delta$. It is well-known that F is absolutely continuous if and only if there exists a Lebesgue integrable map $f : [a, b] \to \mathbb{R}$ such that

$$F(t) = F(a) + \int_{[a,t]} f \, \mathrm{d}\mathfrak{m},$$

for all $t \in [a, b]$, where $\mathfrak{m}$ denotes the Lebesgue measure. Moreover, F is differentiable at $\mathfrak{m}$-almost every point of $[a, b]$ and $F' = f$ $\mathfrak{m}$-almost everywhere. A map F of class C^1, a (continuous) map F that is piecewise C^1 or an everywhere differentiable map F whose derivative is Lebesgue integrable are all examples of absolutely continuous maps. A function $F : \mathbb{R} \to \mathbb{R}$ whose restriction to every compact interval $[a, b]$ is absolutely continuous is called *locally absolutely continuous*.

Example 8.3. Let X be a random variable, i.e., X is a random object taking values on the real line $\mathbb{R}$ endowed with the Borel σ-algebra. Let $\mathfrak{m}$ be the Lebesgue measure restricted to the Borel σ-algebra of $\mathbb{R}$. It follows directly from the facts discussed above that $\mathbb{P}_X$ is absolutely continuous with respect to $\mathfrak{m}$ if and only if the cumulative distribution function $F_X : \mathbb{R} \to \mathbb{R}$ of X is locally absolutely continuous. Moreover, if F_X is locally absolutely continuous then any nonnegative Borel-measurable map $f_X : \mathbb{R} \to [0, +\infty[$ that is $\mathfrak{m}$-almost everywhere equal to the derivative of F_X is a probability density function of X with respect to $\mathfrak{m}$.

Unless otherwise stated, probability density functions for random variables will always be considered with respect to the Lebesgue measure $\mathfrak{m}$.

Example 8.4. Let $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ be measurable spaces and let

$$\mu : \mathcal{B} \longrightarrow [0, +\infty], \quad \nu : \mathcal{B}' \longrightarrow [0, +\infty]$$

be nonnegative countably additive σ-finite measures. We denote by

$$\mu \otimes \nu : \mathcal{B} \otimes \mathcal{B}' \longrightarrow [0, +\infty]$$

the product measure, which is the unique nonnegative countably additive measure such that $(\mu \otimes \nu)(B \times B') = \mu(B)\nu(B')$, for all $B \in \mathcal{B}$ and all $B' \in \mathcal{B}'$. Let X be an $(M, \mathcal{B})$-valued random object on a probability space and Y be an $(M', \mathcal{B}')$-valued random object on that same probability space. Assume that the joint distribution $\mathbb{P}_{(X,Y)}$ is absolutely continuous with respect to $\mu \otimes \nu$, i.e., that there exists a probability density function

$$f_{(X,Y)} : M \times M' \longrightarrow [0, +\infty[$$

of the random object (X, Y) with respect to $\mu \otimes \nu$. The map $f_{(X,Y)}$ is usually called a *joint probability density function* for X and Y. It follows from Fubini–Tonelli's Theorem that the function f_X defined by

$$f_X(x) = \int_{M'} f_{(X,Y)}(x, y)\, \mathrm{d}\nu(y), \quad x \in M$$

is measurable, μ-almost everywhere finite and that

$$\int_B f_X \, \mathrm{d}\mu = \int_{B \times M'} f_{(X,Y)} \, \mathrm{d}(\mu \otimes \nu) = \mathbb{P}_{(X,Y)}(B \times M') = \mathbb{P}_X(B),$$

for all $B \in \mathcal{B}$. Replacing any infinite values that f_X might attain on a set of μ-measure zero with some fixed finite value, we obtain a probability density function for X with respect to μ. Similarly, the map

$$f_Y(y) = \int_M f_{(X,Y)}(x, y)\, \mathrm{d}\mu(x), \quad y \in M'$$

becomes a probability density function for Y with respect to ν after infinite values attained on a set of ν-measure zero are replaced with some fixed finite value.

Thus, probability density functions for the (marginal) distributions of X and Y can be obtained by integrating away the undesired variable from the joint probability density function. Note that this observation generalizes equalities (6.1) that were obtained in the case of discrete random objects.

As we have seen above, the existence of a joint probability density function for (X, Y) with respect to $\mu \otimes \nu$ implies the existence of probability density functions for X and Y with respect to μ and ν, but the converse is not true. For an extreme example, let X and Y be random variables satisfying some functional relation $Y = g(X)$, with $g : \mathbb{R} \to \mathbb{R}$ a measurable function. In this case the $\mathbb{P}_{(X,Y)}$-probability of the graph of g is equal to 1, yet the Lebesgue measure of such graph is zero, so that $\mathbb{P}_{(X,Y)}$ is never absolutely continuous with respect to the Lebesgue measure of $\mathbb{R}^2$.

Example 8.5. One can easily generalize Example 8.4 to arbitrary n-tuples $(X_1, \ldots, X_n)$ of random objects and of nonnegative countably additive σ-finite measures $(\mu_1, \ldots, \mu_n)$, with X_i taking values in a measurable space $(M_i, \mathcal{B}_i)$ and μ_i defined on $\mathcal{B}_i$, for $i = 1, \ldots, n$. The most important particular case happens when $M_i = \mathbb{R}$, $\mathcal{B}_i$ is the Borel σ-algebra and $\mu_i = \mathfrak{m}$ is the restriction of the Lebesgue measure to the Borel σ-algebra. In this case the product measure $\bigotimes_{i=1}^n \mu_i$ is just the Lebesgue measure of $\mathbb{R}^n$ restricted to the Borel σ-algebra of $\mathbb{R}^n$. We denote such measure also by $\mathfrak{m}$ and, unless otherwise stated, the probability density of an $\mathbb{R}^n$-valued random vector $(X_1, \ldots, X_n)$ will be always considered with respect to $\mathfrak{m}$. In elementary probability theory textbooks, what is usually meant by "probability density function" is the probability density function of a random variable or of an $\mathbb{R}^n$-valued random vector with respect to the Lebesgue measure.

Example 8.6. Let $(M, \mathcal{B})$ be a measurable space and let μ and ν be nonnegative countably additive σ-finite measures on $\mathcal{B}$ such that ν is absolutely continuous with respect to μ. If X is an $(M, \mathcal{B})$-valued random object such that $\mathbb{P}_X$ is absolutely continuous with respect to ν then $\mathbb{P}_X$ is also absolutely continuous with respect to μ. Moreover, if f_X is a probability density function for X with respect to ν then a probability density function for X with respect to μ is obtained by multiplying f_X by a Radon–Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ of ν with respect to μ.

Example 8.7. Let $\phi : M \to M'$ be an *isomorphism* between measurable spaces $(M, \mathcal{B})$ and $(M', \mathcal{B}')$, i.e., ϕ is a bijective measurable map whose inverse is also measurable. Let $\mu : \mathcal{B} \to [0, +\infty]$ be a nonnegative countably additive σ-finite measure and X be an $(M, \mathcal{B})$-valued random object such that $\mathbb{P}_X$ is absolutely continuous with respect to μ. We have then that $\mathbb{P}_{\phi(X)} = \phi_* \mathbb{P}_X$ is absolutely continuous with respect to $\phi_* \mu : \mathcal{B}' \to [0, +\infty]$ and, moreover, if f_X is a probability density function for X with respect to μ then $f_X \circ \phi^{-1}$ is a probability density function for $\phi(X)$ with respect to $\phi_* \mu$. The latter statement follows directly from the abstract "change of variables" result for integration with respect to push-forward measures that we will state in Section 9 (Proposition 9.1). Note that if $\phi_* \mu$ is absolutely continuous with respect to some nonnegative countably additive σ-finite measure $\nu : \mathcal{B}' \to [0, +\infty]$ then, as in Example 8.6, a probability density function for $\phi(X)$ with respect to ν is obtained by multiplying $f_X \circ \phi^{-1}$ by a Radon–Nikodym derivative of $\phi_* \mu$ with respect to ν.

Example 8.8. Given $a \in \mathbb{R}$, $b \in \mathbb{R}^n$ with $a \neq 0$, we obviously have that the map $\phi : \mathbb{R}^n \to \mathbb{R}^n$ given by $\phi(x) = ax + b$, for all $x \in \mathbb{R}^n$, is an isomorphism of the measurable space $\mathbb{R}^n$ with itself, where $\mathbb{R}^n$ is endowed with its Borel σ-algebra. Moreover, if $\mathfrak{m}$ denotes the restriction of the Lebesgue measure of $\mathbb{R}^n$ to the Borel σ-algebra then $\phi_* \mathfrak{m} = |a|^{-n} \mathfrak{m}$, so that a Radon–Nikodym derivative of $\phi_* \mathfrak{m}$ with respect to $\mathfrak{m}$ is the function that is constant and equal to $|a|^{-n}$. It then follows from the results stated in Example 8.7 that if X is an $\mathbb{R}^n$-valued random vector that admits a probability density function f_X with respect to $\mathfrak{m}$ then a probability density function for $\phi(X) = aX + b$ with respect to $\mathfrak{m}$ is given by

$$f_{aX+b}(y) = \frac{1}{|a|^n} f_X\Big(\frac{y - b}{a}\Big),$$

for all $y \in \mathbb{R}^n$.

9. Expected value

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable. The *expected value* of X, denoted $E(X)$, is defined by:

$$E(X) = \int_\Omega X \, \mathrm{d}\mathbb{P},$$

if such integral exists. Note that if X is nonnegative, such integral always exists, but possibly takes the value $+\infty$. In general, X has a positive part X^+ and a negative part X^- and the integral of X exists (possibly taking the values $+\infty$ or $-\infty$) if and only if either X^+ or X^- has a finite integral.

We recall the following simple yet useful "change of variables" result for integration with respect to push-forward measures.

Proposition 9.1. *Let $(\Omega, \mathcal{A})$, $(M, \mathcal{B})$ be measurable spaces, $X : \Omega \to M$ be a measurable map, μ be a nonnegative countably additive measure on $\mathcal{A}$ and denote by $X_*\mu : \mathcal{B} \to [0, +\infty]$ the push-forward measure. For every measurable map $f : M \to [-\infty, +\infty]$ we have that the equality*

$$\int_\Omega f \circ X \, \mathrm{d}\mu = \int_M f \, \mathrm{d}(X_*\mu)$$

holds, meaning that the integral on the lefthand side of the equality exists if and only if the integral on the righthand side of the equality exists, with such integrals being equal when both exist. □

Let $X : \Omega \to M$ be a random object taking values in some measurable space $(M, \mathcal{B})$ and let $f : M \to \mathbb{R}$ be a measurable function, so that $f(X)$ is a random variable. Since the distribution $\mathbb{P}_X$ of X is simply the push-forward of the probability measure $\mathbb{P}$ under the map X, Proposition 9.1 yields

$$\int_\Omega f(X) \, \mathrm{d}\mathbb{P} = \int_M f \, \mathrm{d}\mathbb{P}_X,$$

meaning that the integral on the lefthand side of the equality exists if and only if the integral on the righthand side of the equality exists and that they are equal when both exist. Thus

$$E\big(f(X)\big) = \int_M f \, \mathrm{d}\mathbb{P}_X \tag{9.1}$$

and, in particular, if X is a random variable and f is the identity function of $\mathbb{R}$, we get:

$$E(X) = \int_{\mathbb{R}} x \, \mathrm{d}\mathbb{P}_X(x). \tag{9.2}$$

Equality (9.2) can be interpreted as saying that $E(X)$ is the average of the values taken by X weighted by their probabilities. This holds literally if X is discrete, i.e., if the image of X is countable then (9.2) becomes

$$E(X) = \sum_{x \in \mathrm{Im}(X)} x \, \mathbb{P}(X = x)$$

and (9.1) becomes (assuming $\{x\} \in \mathcal{B}$ for all $x \in \mathrm{Im}(X)$):

$$E\big(f(X)\big) = \sum_{x \in \mathrm{Im}(X)} f(x) \, \mathbb{P}(X = x).$$

Note that the name "expected value" is somewhat misleading, as $E(X)$ is not a "value that is expected" in the sense that it has a large probability

of being observed in some sense. It is really just an average and, in fact, if X is discrete it often happens that $E(X)$ is not even in the image of X, so that $E(X)$ is an impossible value for X.

Example 9.2. Let X be a random object taking values in a measurable space $(M, \mathcal{B})$ and μ be a nonnegative countably additive σ-finite measure defined on $\mathcal{B}$. Assume that $\mathbb{P}_X$ is absolutely continuous with respect to μ and let $f_X : M \to [0, +\infty[$ be a probability density function for X with respect to μ. It is well-known from basic measure theory that integrating a measurable map $g : M \to \mathbb{R}$ with respect to $\mathbb{P}_X$ is the same as integrating gf_X with respect to μ and therefore

$$E\big(g(X)\big) = \int_M g \,\mathrm{d}\mathbb{P}_X = \int_M g f_X \,\mathrm{d}\mu,$$

i.e., we have

$$E\big(g(X)\big) = \int_M g f_X \,\mathrm{d}\mu$$

for any measurable map $g : M \to \mathbb{R}$ meaning that the integral on the lefthand side of the equality exists if and only if the integral on the righthand side of the equality exists and that they are equal when both exist.

10. Variance and covariance

Let V be a real vector space endowed with an inner product $\langle \cdot, \cdot \rangle$ and let W be a subspace of V. Assume that the orthogonal projection operator $P : V \to W$ is well-defined (which happens, for instance, if W is finite-dimensional). For every $v \in V$, the point $P(v)$ is the element of W closest to v and $\|v - P(v)\|^2 = \langle v - P(v), v - P(v) \rangle$ is the square of the distance between v and the set W. If we define

$$\langle\!\langle v_1, v_2 \rangle\!\rangle = \langle v_1 - P(v_1), v_2 - P(v_2) \rangle,$$

for all $v_1, v_2 \in V$, then $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is a positive semi-definite symmetric bilinear form on V and $\langle\!\langle v, v \rangle\!\rangle^{\frac{1}{2}}$ is the distance between v and W for all $v \in V$. In other words, the semi-norm induced by $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ gives the distance from W.

Now let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and consider the Hilbert space $L^2(\Omega, \mathcal{A}, \mathbb{P})$ of (equivalence classes of $\mathbb{P}$-almost everywhere equal) square integrable measurable maps $X : \Omega \to \mathbb{R}$ endowed with the inner product:

$$\langle X, Y \rangle = \int_\Omega XY \,\mathrm{d}\mathbb{P} = E(XY), \quad X, Y \in L^2(\Omega, \mathcal{A}, \mathbb{P}).$$

We apply the construction above with $V = L^2(\Omega, \mathcal{A}, \mathbb{P})$ and W the one-dimensional subspace of V consisting of $\mathbb{P}$-almost everywhere constant maps. The orthogonal projection $P : V \to W$ is easily seen to be given by

$$P(X) = E(X),$$

for all $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$, in which we identify the real number $E(X)$ with the map defined on Ω that is constant and equal to $E(X)$. Note that this gives

another nice interpretation for the expected value of X: it is the constant random variable that is L^2-closest to the random variable X.

The positive semi-definite symmetric bilinear form $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ obtained from $\langle \cdot, \cdot \rangle$ and the orthogonal projection $P : V \to W$ will be called the covariance map. More explicitly, we give the following definition.

Definition 10.1. For all $X, Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$, we define the *covariance* of X and Y by setting:

$$\operatorname{Cov}(X, Y) = \langle X - E(X), Y - E(Y) \rangle = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big].$$

A simple computation yields:

$$\operatorname{Cov}(X, Y) = E(XY) - E(X)E(Y),$$

for all $X, Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. The square of the semi-norm induced by the covariance map is called the variance map.

Definition 10.2. For all X in $L^2(\Omega, \mathcal{A}, \mathbb{P})$, we define its *variance* by setting:

$$\operatorname{Var}(X) = \operatorname{Cov}(X, X) = E\big[\big(X - E(X)\big)^2\big] = E(X^2) - E(X)^2.$$

The equality $\operatorname{Var}(X) = E\big[\big(X - E(X)\big)^2\big]$ can also be used to define the variance of X in case X is not square integrable, but in this case $\operatorname{Var}(X)$ is always equal to $+\infty$.

We thus have that the variance $\operatorname{Var}(X)$ is the squared L^2-distance between X and the one-dimensional space of $\mathbb{P}$-almost everywhere constant functions. In particular, $\operatorname{Var}(X) = 0$ if and only if X is $\mathbb{P}$-almost everywhere constant or, in the probability theoretic jargon, $\operatorname{Var}(X) = 0$ if and only if X is almost surely constant. The variance of X can be seen as a measure of how spread out the distribution of X is on the real line. A small variance $\operatorname{Var}(X)$ means that the values of X tend to fall near to the expected value $E(X)$.

Applying the Cauchy–Schwarz inequality to the positive semi-definite symmetric bilinear form Cov we obtain

$$|\operatorname{Cov}(X, Y)| \leq \operatorname{Var}(X, X)^{\frac{1}{2}} \operatorname{Var}(Y, Y)^{\frac{1}{2}}, \tag{10.1}$$

for all $X, Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. For inner products (i.e., positive definite symmetric bilinear forms) the equality holds in the Cauchy–Schwarz inequality if and only if the vectors are linearly dependent, but for positive semi-definite symmetric bilinear forms it holds if and only if the vectors are linearly dependent modulo the kernel of the bilinear form (which for Cov consists of almost surely constant maps). Hence equality holds in (10.1) if and only if either X is almost surely constant or $Y = aX + b$ almost surely for certain $a, b \in \mathbb{R}$.

If $X, Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ and neither X nor Y is almost surely constant, we define the *correlation* between X and Y by setting:

$$\rho(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(X, X)^{\frac{1}{2}} \operatorname{Var}(Y, Y)^{\frac{1}{2}}}.$$

By the Cauchy–Schwarz inequality (10.1), we have $|\rho(X,Y)| \leq 1$. Moreover, $\rho(X,Y) = 1$ if and only if $Y = aX + b$ almost surely for certain $a, b \in \mathbb{R}$ with $a > 0$ and $\rho(X,Y) = -1$ if and only if $Y = aX + b$ for certain $a, b \in \mathbb{R}$ with $a < 0$.

The square root $\mathrm{Var}(X)^{\frac{1}{2}}$ of the variance of X — that is, the semi-norm of X induced by the covariance map Cov — is usually called the *standard deviation* of X and it is equal to the L^2-distance between X and the one-dimensional space of almost surely constant maps. The correlation $\rho(X,Y)$ can be interpreted geometrically as the cosine of the angle between X and Y with respect to the covariance map.

11. Expectation of random vectors and the covariance matrix

Let V be a real finite-dimensional vector space endowed with its Borel σ-algebra $\mathcal{B}$, where the topology of V is the canonical topology (induced by an arbitrary norm). Let us discuss the notions of expectation, covariance and variance for V-valued random objects. Recall that a V-valued random object is also called a V-valued random vector.

The theory of integration with respect to a measure for V-valued measurable functions is a simple extension of the theory of integration of real-valued measurable functions. Here we focus on integration with respect to a probability measure. Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a random vector $X : \Omega \to V$ there exists a unique vector $\int_\Omega X \,\mathrm{d}\mathbb{P} \in V$ such that

$$\alpha\left(\int_\Omega X \,\mathrm{d}\mathbb{P}\right) = \int_\Omega \alpha \circ X \,\mathrm{d}\mathbb{P},$$

for every linear functional $\alpha \in V^*$, provided that $\alpha(X) = \alpha \circ X$ has a finite integral with respect to $\mathbb{P}$ for every $\alpha \in V^*$. Here, as usual, V^* denotes the dual space of V. This fact is easily proven using a basis of V and defining the integral $\int_\Omega X \,\mathrm{d}\mathbb{P}$ coordinatewise.

Note that, for any $p \in [1, +\infty[$, the following statements about a random vector $X : \Omega \to V$ are equivalent:

(i) $|\alpha(X)|^p$ has finite integral with respect to $\mathbb{P}$ for all $\alpha \in V^*$;

(ii) V^* is contained in $L^p(V, \mathcal{B}, \mathbb{P}_X)$, i.e., the map $V \ni v \mapsto |\alpha(v)|^p$ has finite integral with respect to $\mathbb{P}_X$ for every $\alpha \in V^*$;

(iii) the map $\Omega \ni \omega \mapsto \|X(\omega)\|^p$ has finite integral, where $\|\cdot\|$ is some fixed arbitrary norm in V.

When any of these conditions is satisfied, we will say that X is *p-th integrable*; for $p = 1$ we simply say that X is *integrable* and for $p = 2$ we say that X is *square integrable*. For any $p \in [1, +\infty[$, we denote by $L^p(\Omega, \mathcal{A}, \mathbb{P}; V)$ the space of (equivalence classes of $\mathbb{P}$-almost everywhere equal) p-th integrable measurable maps $X : \Omega \to V$.

Definition 11.1. For an integrable random vector $X : \Omega \to V$, its *expected value* is defined by:

$$E(X) = \int_\Omega X \, \mathrm{d}\mathbb{P} \in V.$$

Clearly, for any linear transformation $T : V \to W$ taking values in some other real finite-dimensional vector space W we have:

$$E\big(T(X)\big) = \int_\Omega T \circ X \, \mathrm{d}\mathbb{P} = T\left(\int_\Omega X \, \mathrm{d}\mathbb{P} \right) = T\big(E(X)\big). \tag{11.1}$$

Remark 11.2. If $X : \Omega \to V$ is a V-valued random vector and W is a subspace of V containing the image of X then we can regard X also as a W-valued random vector. Using (11.1) with $T : W \to V$ the inclusion map, we see that X has the same expected value when regarded as a V-valued random vector and as a W-valued random vector. In particular, if the image of X is contained in a subspace W then $E(X) \in W$. Recall that an *affine subspace* of a vector space V is a translation $v + W = \big\{v + w : w \in W\big\}$ of some vector subspace W of V. Note that if the image of X is contained in an affine subspace $v + W$ of V then $E(X)$ is in $v + W$, since $E(X - v)$ is in W and $E(v) = v$ for any $v \in V$ (regarded as a constant V-valued random vector).

How do we go about defining variance and covariance for square integrable random vectors? To motivate the definitions we will think in terms of tensor products. The official definitions presented after the motivation will not be dependent on the facts used in the reasoning below.

Observe first that there is a natural identification between $L^p(\Omega, \mathcal{A}, \mathbb{P}; V)$ and the tensor product $L^p(\Omega, \mathcal{A}, \mathbb{P}) \otimes V$ given by the isomorphism:

$$L^p(\Omega, \mathcal{A}, \mathbb{P}) \otimes V \ni X \otimes v \longmapsto Xv \in L^p(\Omega, \mathcal{A}, \mathbb{P}; V).$$

Moreover, the covariance map of square integrable random variables is a bilinear map

$$\mathrm{Cov} : L^2(\Omega, \mathcal{A}, \mathbb{P}) \times L^2(\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow \mathbb{R}$$

and for real finite-dimensional vector spaces V and W such map naturally induces a bilinear map

$$\big(L^2(\Omega, \mathcal{A}, \mathbb{P}) \otimes V\big) \times \big(L^2(\Omega, \mathcal{A}, \mathbb{P}) \otimes W\big) \longrightarrow V \otimes W$$

that sends $(X \otimes v, Y \otimes w)$ to $\mathrm{Cov}(X, Y)(v \otimes w)$, for all $X, Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$, all $v \in V$ and all $w \in W$. Using the identifications

$$L^2(\Omega, \mathcal{A}, \mathbb{P}; V) \cong L^2(\Omega, \mathcal{A}, \mathbb{P}) \otimes V, \quad L^2(\Omega, \mathcal{A}, \mathbb{P}; W) \cong L^2(\Omega, \mathcal{A}, \mathbb{P}) \otimes W$$

we then obtain a bilinear map

$$\mathrm{Cov} : L^2(\Omega, \mathcal{A}, \mathbb{P}; V) \times L^2(\Omega, \mathcal{A}, \mathbb{P}; W) \longrightarrow V \otimes W$$

that will be called the covariance map for V-valued and W-valued random vectors.

Recall that, since V and W are finite-dimensional, the tensor product $V \otimes W$ can be naturally identified with the space of bilinear maps from $V^* \times W^*$ to $\mathbb{R}$ by setting

$$(v \otimes w)(\alpha, \beta) = \alpha(v)\beta(w),$$

for all $\alpha \in V^*$ and all $\beta \in W^*$. From now on we will use such identification throughout, i.e., we will simply regard the tensor product $V \otimes W$ as a notation for such space of bilinear maps.

Here is our official definition of covariance of square integrable random vectors. Such definition is equivalent to what was described above.

Definition 11.3. Given real finite-dimensional vector spaces V and W, a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and square integrable random vectors $X : \Omega \to V$ and $Y : \Omega \to W$, we define their *covariance* $\mathrm{Cov}(X, Y) \in V \otimes W$ by setting:

$$\mathrm{Cov}(X, Y)(\alpha, \beta) = \mathrm{Cov}\big(\alpha(X), \beta(Y)\big),$$

for all $\alpha \in V^*$ and all $\beta \in W^*$. The *variance* of the square integrable random vector $X : \Omega \to V$ is defined by:

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) \in V \otimes V.$$

We have

$$\mathrm{Var}(X)(\alpha, \beta) = \mathrm{Cov}\big(\alpha(X), \beta(X)\big),$$

for all $\alpha, \beta \in V^*$, so that $\mathrm{Var}(X)$ is a positive semi-definite symmetric bilinear form on V^*. Since $\mathrm{Var}(X)$ is positive semi-definite, its *kernel*, i.e., the subspace of V^* given by

$$\mathrm{Ker}\big(\mathrm{Var}(X)\big) = \big\{\alpha \in V^* : \mathrm{Var}(X)(\alpha, \beta) = 0 \text{ for all } \beta \in V^*\big\}$$

coincides with the set of those $\alpha \in V^*$ with $\mathrm{Var}(X)(\alpha, \alpha) = \mathrm{Var}\big(\alpha(X)\big) = 0$. Hence:

$$\mathrm{Ker}\big(\mathrm{Var}(X)\big) = \big\{\alpha \in V^* : \alpha(X) \text{ is almost surely constant}\big\}. \tag{11.2}$$

Using this equality we can show that $\mathrm{Var}(X)$ is *degenerate*, i.e., has a nonzero kernel, if and only if the support of the distribution $\mathbb{P}_X$ of X is contained in a proper affine subspace of V. The definition of support of a measure on a topological space is recalled below.

Definition 11.4. If M is a topological space and μ is a nonnegative countably additive measure defined on the Borel σ-algebra of M then the *support* of μ is defined as the complement in M of the union of all open subsets of M having zero measure, in case such union also has zero measure. This is always the case if the topology of M is second countable, as in this case such union can be replaced with a countable union.

Thus, saying that the support of a probability measure on the Borel σ-algebra of a second countable topological space is contained in a closed subset simply means that such closed subset has probability equal to one.

Proposition 11.5. *Let $X : \Omega \to V$ be a square integrable random vector on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where V is a real finite-dimensional vector space. If*

$$W = \big\{ v \in V : \alpha(v) = 0 \text{ for all } \alpha \in \operatorname{Ker}\big(\operatorname{Var}(X)\big) \big\}$$

denotes the subspace of V annihilated by $\operatorname{Ker}\big(\operatorname{Var}(X)\big)$ then the support of $\mathbb{P}_X$ is contained in some translation $v+W$ of W. Moreover, no proper affine subspace of $v + W$ contains the support of $\mathbb{P}_X$.

Proof. If $(\alpha_i)_{i=1}^k$ is a basis for $\operatorname{Ker}\big(\operatorname{Var}(X)\big)$ then

$$T : V \ni v \longmapsto \big(\alpha_1(v), \ldots, \alpha_k(v)\big) \in \mathbb{R}^k$$

is a linear map whose kernel is W and, by (11.2), the random vector $T(X)$ is almost surely equal to some constant $T(v)$, for some $v \in V$. Hence the support of $\mathbb{P}_X$ is contained in $v + W$.

If the support of $\mathbb{P}_X$ were contained in some proper affine subspace of $v + W$ then such affine subspace would be a translation of a proper vector subspace W' of W. This would imply that every $\alpha \in V^*$ that annihilates W' is such that $\alpha(X)$ is almost surely constant and thus that the annihilator of W' is contained in $\operatorname{Ker}\big(\operatorname{Var}(X)\big)$. But this is not possible, as the annihilator of W' properly contains the annihilator of W, which is equal to $\operatorname{Ker}\big(\operatorname{Var}(X)\big)$. □

Corollary 11.6. *Let $X : \Omega \to V$ be a square integrable random vector on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where V is a real finite-dimensional vector space. We have that the symmetric bilinear form $\operatorname{Var}(X)$ is nondegenerate (and thus positive definite) if and only if there is no proper affine subspace of V containing the support of $\mathbb{P}_X$.* □

11.1. **Matrix representation of the variance.** If $(e_i)_{i=1}^n$ is a basis of V, then the symmetric bilinear form $\operatorname{Var}(X)$ on V^* is represented by an $n \times n$ symmetric matrix with respect to the dual basis $(\alpha_i)_{i=1}^n$ of $(e_i)_{i=1}^n$. The element on the i-th row and j-th column of that matrix is equal to

$$\operatorname{Var}(X)(\alpha_i, \alpha_j) = \operatorname{Cov}\big(\alpha_i(X), \alpha_j(X)\big),$$

for all $i, j = 1, \ldots, n$. In a different perspective, we have that $(e_i \otimes e_j)_{i,j=1}^n$ is a basis of $V \otimes V$ and

$$\operatorname{Var}(X) = \sum_{i,j=1}^n \operatorname{Var}(X)(\alpha_i, \alpha_j)\, e_i \otimes e_j = \sum_{i=1}^n \operatorname{Cov}\big(\alpha_i(X), \alpha_j(X)\big)\, e_i \otimes e_j.$$

Definition 11.7. Given a square integrable V-valued random vector X and a basis $(e_i)_{i=1}^n$ of V with dual basis $(\alpha_i)_{i=1}^n$ then the $n \times n$ symmetric matrix whose entry in the i-th row and j-th column is $\operatorname{Cov}\big(\alpha_i(X), \alpha_j(X)\big)$ is called the *covariance matrix* of X with respect to the basis $(e_i)_{i=1}^n$.

Standard textbooks usually focus only on the case that $V = \mathbb{R}^n$. In that case, we have a canonical basis $(e_i)_{i=1}^n$ and X is of the form $(X_1, \ldots, X_n)$

with $X_i = \alpha_i(X)$, for all $i = 1, \dots, n$, where $(\alpha_i)_{i=1}^n$ is dual to the canonical basis. Thus the corresponding covariance matrix is

$$\big(\mathrm{Cov}(X_i, X_j)\big)_{i,j=1}^n.$$

We will call this simply the *covariance matrix* of the $\mathbb{R}^n$-valued random vector X, without explicit mention to the canonical basis.

11.2. **Naturality of covariance with respect to linear maps.** A pair of linear maps $T_1 : V_1 \to W_1$, $T_2 : V_2 \to W_2$ induces a linear map

$$T_1 \otimes T_2 : V_1 \otimes V_2 \ni v_1 \otimes v_2 \longmapsto T_1(v_1) \otimes T_2(v_2) \in W_1 \otimes W_2$$

between tensor products. Identifying as before elements of $V_1 \otimes V_2$ with bilinear maps $B : V_1^* \times V_2^* \to \mathbb{R}$, we have:

$$(T_1 \otimes T_2)(B) = B(T_1^* \cdot, T_2^* \cdot),$$

or, more explicitly

$$(T_1 \otimes T_2)(B)(\beta_1, \beta_2) = B\big(T_1^*(\beta_1), T_2^*(\beta_2)\big) = B(\beta_1 \circ T_1, \beta_2 \circ T_2),$$

for all $\beta_1 \in V_1^*$ and all $\beta_2 \in V_2^*$. The following result follows directly from the definitions.

Proposition 11.8. *Let $T_1 : V_1 \to W_1$ and $T_2 : V_2 \to W_2$ be linear maps between real finite-dimensional vector spaces, X be a square integrable V_1-valued random vector on a probability space and Y be a square integrable V_2-valued random vector on that same probability space. We have:*

$$\mathrm{Cov}\big(T_1(X), T_2(Y)\big) = (T_1 \otimes T_2)\big(\mathrm{Cov}(X, Y)\big). \qquad \square$$

Corollary 11.9. *Let $T : V \to W$ be a linear map between real finite-dimensional vector spaces and X be a V-valued square integrable random vector. We have:*

$$\mathrm{Var}\big(T(X)\big) = (T \otimes T)\big(\mathrm{Var}(X)\big). \qquad \square$$

11.3. **Random objects with values in an abstract affine space.** The theory developed in this section can be readily generalized to random objects that take values in a real finite-dimensional abstract affine space instead of a vector space. Such generalization is sometimes convenient (see Proposition 11.5).

Recall that an *affine space* is a nonempty set P endowed with a transitive action

$$P \times V \ni (p, v) \longmapsto p + v \in P$$

without fixed points of the additive group of a vector space V. For $p, q \in P$, we write $p - q$ for the unique vector in V such that $q + (p - q) = p$. Each choice of a point $O \in P$ — usually called an *origin* — leads to an identification $V \ni v \mapsto O + v \in P$ between V and P. A distinct choice of origin leads to a different identification between V and P that differs from the first by a translation of V. We call V the vector space *parallel* to P.

We assume below that V is real and finite-dimensional, so that it has a canonical topology which induces a canonical topology on P through any choice of an origin. Let P be endowed with the corresponding Borel σ-algebra.

Given a random object X with values in P, we can define the *expected value* of X by setting

$$E(X) = E(X - O) + O \in P,$$

for any choice of $O \in P$, provided that $\omega \mapsto (X - O)(\omega) = X(\omega) - O \in V$ is integrable. It is easily checked that $E(X)$ does not depend on the choice of O. The fact that the measure is a probability measure is crucial here!

Covariance and variance can also be defined for random objects taking values in affine spaces. If P is an affine space parallel to a real finite-dimensional vector space V and Q is an affine space parallel to a real finite-dimensional vector space W, we define

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X - O, Y - O') \in V \otimes W,$$

for a P-valued random object X and a Q-valued random object Y such that $X - O$ and $Y - O'$ are both square integrable, where $O \in P$ and $O' \in Q$ are chosen arbitrarily. The definition of $\mathrm{Cov}(X, Y)$ does not depend on the choices of O and O'. The variance of X is then defined by:

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) \in V \otimes V.$$

There are obvious generalizations of Proposition 11.5, Corollary 11.6, Proposition 11.8 and Corollary 11.9 to the context of affine space-valued random objects.

12. Convergence of random variables

In measure theory courses one studies several notions of convergence for real-valued measurable functions defined on a measure space. Since random variables are also real-valued measurable functions on a measure space, all such notions of convergence can be used for random variables. Probabilists have their own favorite names for such convergence notions and below we present the suitable translations from measure theory language to probability theory language.

Let $(X_n)_{n\geq 1}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and let X be another random variable on that same probability space. In probability theory we will say that $(X_n)_{n\geq 1}$ *converges almost surely* to X if $(X_n)_{n\geq 1}$ converges pointwise almost everywhere to X, i.e., if

$$\lim_{n\to+\infty} X_n(\omega) = X(\omega)$$

for all $\omega \in \Omega$ outside of some measurable subset of Ω with zero probability. We say that $(X_n)_{n\geq 1}$ *converges in probability* to X if $(X_n)_{n\geq 1}$ converges in measure to X, i.e., if for all $\varepsilon > 0$ we have:

$$\lim_{n\to+\infty} \mathbb{P}\big(|X_n - X| \geq \varepsilon\big) = 0.$$

Recall that a sequence of real-valued measurable functions on a measure space is said to converge *almost uniformly* to another real-valued measurable function if for every $\varepsilon > 0$ there exists a measurable subset with measure less than ε outside of which the convergence is uniform. Due to Egoroff's Theorem, for finite measure spaces, almost uniform convergence is equivalent to pointwise convergence almost everywhere. Since probability measures are finite, almost uniform convergence does not give rise to a new notion of convergence of random variables, i.e., it is equivalent to convergence almost surely. Note that since almost uniform convergence implies convergence in measure, we have that almost surely convergence of random variables implies convergence in probability. Standard results from measure theory courses also yield that if $(X_n)_{n\geq 1}$ converges to X in probability then some subsequence of $(X_n)_{n\geq 1}$ converges to X almost surely.

Another important notion of convergence for real-valued measurable functions is convergence with respect to the L^p-norm for some $p \in [1, +\infty[$. In the context of probability theory, this is called convergence in the p-th mean or simply *convergence in the mean* if $p = 1$. Thus, $(X_n)_{n\geq 1}$ *converges in the p-th mean* to X if and only if:

$$\lim_{n\to+\infty} E\big(|X_n - X|^p\big) = 0.$$

Clearly, convergence in the p-th mean is stronger than convergence in probability. Moreover, since the probability measure is finite, for $1 \leq q \leq p$, convergence in the p-th mean implies convergence in the q-th mean. By the Dominated Convergence Theorem, if $|X_n| \leq |Y|$ almost surely for all $n \geq 1$ and some $Y \in L^p(\Omega, \mathcal{A}, \mathbb{P})$, then almost surely convergence of $(X_n)_{n\geq 1}$ implies convergence in the p-th mean.

We prove below a couple of other results regarding convergence of random variables that are less well-known from measure theory courses.

Proposition 12.1. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables on a probability space and X be a random variable on that same probability space. If $f : D \to \mathbb{R}$ is a continuous function defined in a subset D of $\mathbb{R}$ containing the image of X and the image of all X_n and if $(X_n)_{n\geq 1}$ converges in probability to X then $\big(f(X_n)\big)_{n\geq 1}$ converges in probability to $f(X)$.*

Proof. Assuming by contradiction that the thesis is false, one obtains $\varepsilon > 0$, $\eta > 0$ and a strictly increasing sequence $(n_k)_{k\geq 1}$ of positive integers with

$$\mathbb{P}\big(|f(X_{n_k}) - f(X)| \geq \varepsilon\big) \geq \eta, \tag{12.1}$$

for all $k \geq 1$. Passing to a subsequence, we may assume that $(X_{n_k})_{k\geq 1}$ converges to X almost surely, which yields that $\big(f(X_{n_k})\big)_{k\geq 1}$ converges almost surely to $f(X)$. But almost surely convergence implies convergence in probability and this yields a contradiction with (12.1). □

Proposition 12.2. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables on a probability space and X be a random variable on that same probability space.*

If $(X_n)_{n\geq 1}$ converges in probability to X and $|X_n| \leq Y$ almost surely for all $n \geq 1$ and some $Y \in L^p(\Omega, \mathcal{A}, \mathbb{P})$ then $(X_n)_{n\geq 1}$ converges to X in the p-th mean and in particular $\lim_{n\to+\infty} E(X_n) = E(X)$.

Proof. Similar to the proof of Proposition 12.1: assume by contradiction that the thesis is false, pick a subsequence that converges almost surely and apply the Dominated Convergence Theorem. □

We note that the definitions and results discussed in this section are easily generalized to random objects taking values in a separable metric space (M, d) endowed with its Borel σ-algebra. Separability of (M, d) is required for instance because if X and Y are random objects taking values in M we will often need the function $d(X, Y) : \Omega \to \mathbb{R}$ to be measurable and this requires separability (see Remark 6.7).

13. Topologies for the set of probability measures

The notions of convergence of random variables discussed in Section 12 can all be seen as stating that the random variable X_n "becomes close" to the random variable X as n goes to $+\infty$. This should imply that also the distribution of X_n becomes close to the distribution of X in some sense. However, in some situations, we are just interested in the closeness of the distributions and we do not care about the closeness of the random variables. We then need a notion of convergence for probability measures.

Let $(M, \mathcal{B})$ be a measurable space and denote by $\mathrm{Prob}(M, \mathcal{B})$ the set of all probability measures defined on $\mathcal{B}$. Let us discuss some possible topologies for the set $\mathrm{Prob}(M, \mathcal{B})$. Such topologies will, of course, correspond to notions of convergence of probability measures. It is well known that the space $\mathrm{ca}(M, \mathcal{B})$ of all finite signed countably additive measures defined on $\mathcal{B}$ is a Banach space endowed with the total variation norm $\|\mu\| = |\mu|(M)$. We explain below why the topology induced by such norm is usually not a very useful topology for $\mathrm{Prob}(M, \mathcal{B})$. We need a definition.

Definition 13.1. Given a measurable space $(M, \mathcal{B})$, for each $x \in M$, we denote by $\delta_x : \mathcal{B} \to [0, 1]$ the probability measure defined by $\delta_x(B) = 1$ if $x \in B$ and $\delta_x(B) = 0$ otherwise. This is called the *Dirac delta* probability measure centered at x.

We assume that all singletons $\{x\}$ with $x \in M$ are in $\mathcal{B}$ to avoid pathologies. The Dirac delta probability measure δ_x models a degenerate random experiment in which the outcome x is obtained with certainty. Note that if $x, y \in M$ are distinct, the distance $\|\delta_x - \delta_y\|$ between δ_x and δ_y with respect to the total variation norm is equal to 2.

Now assume that M is endowed with some topology (say, M is a metric space) and $\mathcal{B}$ is the Borel σ-algebra. In the context of real-world applications, if x and y are very very close, the probability measures δ_x and δ_y are indistinguishable, as experimental equipment has limited precision. In this context it is thus completely inappropriate that the distance between

δ_x and δ_y remains fixed no matter how close $y \neq x$ becomes to x. Having this example in mind, a good requirement for a topology in $\mathrm{Prob}(M, \mathcal{B})$ is that the map

$$\delta : M \ni x \longmapsto \delta_x \in \mathrm{Prob}(M, \mathcal{B})$$

be continuous. In order to find such a topology, we look for topologies that are weaker than the total variation norm topology.

A first possible candidate is the weak topology of the Banach space $\mathrm{ca}(M, \mathcal{B})$, i.e., the smallest topology that makes all norm-continuous linear functionals continuous. But this topology does not satisfy our requirement: namely, any bounded measurable function $f : M \to \mathbb{R}$ defines a norm-continuous linear functional

$$\mathrm{ca}(M, \mathcal{B}) \ni \mu \longmapsto \int_M f \, \mathrm{d}\mu \in \mathbb{R} \tag{13.1}$$

and the composition of (13.1) with the map δ is simply the function f. Except for trivial cases, it is not true that every bounded measurable function is continuous and thus δ is not continuous with respect to the weak topology of the Banach space $\mathrm{ca}(M, \mathcal{B})$.

The considerations above yield a suggestion of topology for $\mathrm{Prob}(M, \mathcal{B})$. Namely, endow $\mathrm{ca}(M, \mathcal{B})$ and $\mathrm{Prob}(M, \mathcal{B})$ with the smallest topology that makes the linear functional (13.1) continuous for every bounded continuous function $f : M \to \mathbb{R}$. This will obviously make the map δ continuous. In probability theory textbooks this topology is usually called the *weak topology* on $\mathrm{Prob}(M, \mathcal{B})$, but one should be careful to distinguish it from the weak topology of the Banach space $\mathrm{ca}(M, \mathcal{B})$.

Using the terminology that is normally employed in topological vector spaces books, the topology defined above is the weak topology on $\mathrm{ca}(M, \mathcal{B})$ induced by the bilinear pairing

$$\mathrm{ca}(M, \mathcal{B}) \times C_{\mathrm{b}}(M, \mathbb{R}) \ni (\mu, f) \longmapsto \int_M f \, \mathrm{d}\mu \in \mathbb{R}, \tag{13.2}$$

where $C_{\mathrm{b}}(M, \mathbb{R})$ denotes the space of bounded continuous real-valued maps defined on M.

Remark 13.2. If M is metrizable then standard regularity results for finite Borel measures on metric spaces plus the Hahn Decomposition Theorem for signed measures imply that

$$\|\mu\| = \sup_{f \in C_{\mathrm{b}}(M,R)} \int_M f \, \mathrm{d}\mu.$$

It follows that the linear functionals (13.1) with f bounded and continuous separate the points of $\mathrm{ca}(M, \mathcal{B})$ and hence the weak topology corresponding to the bilinear pairing (13.2) is Hausdorff.

Definition 13.3. Let M be a topological space endowed with its Borel σ-algebra $\mathcal{B}$. If $(X_n)_{n \geq 1}$ is a sequence of $(M, \mathcal{B})$-valued random objects and X is an $(M, \mathcal{B})$-valued random object, we say that $(X_n)_{n \geq 1}$ *converges in*

distribution to X if $\lim_{n\to+\infty} \mathbb{P}_{X_n} = \mathbb{P}_X$ with respect to the weak topology of $\mathrm{Prob}(M, \mathcal{B})$ defined above. In other words, $(X_n)_{n\geq 1}$ converges in distribution to X if and only if

$$\lim_{n\to+\infty} E\big(f(X_n)\big) = E\big(f(X)\big), \tag{13.3}$$

for every bounded continuous function $f : M \to \mathbb{R}$.

Note that for the definition above it is not even relevant that the objects X_n and X be all defined on the same probability space.

The following is a direct consequence of the generalizations of Propositions 12.1 and 12.2 to random objects taking values in a separable metric space.

Proposition 13.4. *Let M be a separable metric space endowed with its Borel σ-algebra $\mathcal{B}$. Let $(X_n)_{n\geq 1}$ be a sequence of $(M, \mathcal{B})$-valued random objects on the same probability space and X be an $(M, \mathcal{B})$-valued random object on that same probability space. If $(X_n)_{n\geq 1}$ converges in probability to X then $(X_n)_{n\geq 1}$ converges in distribution to X.* □

The converse of Proposition 13.4 does not hold in general, but it does if X is almost surely constant.

Proposition 13.5. *Let M be a separable metric space endowed with its Borel σ-algebra $\mathcal{B}$. Let $(X_n)_{n\geq 1}$ be a sequence of $(M, \mathcal{B})$-valued random objects on the same probability space and X be an $(M, \mathcal{B})$-valued random object on that same probability space. If X is almost surely constant and $(X_n)_{n\geq 1}$ converges in distribution to X then $(X_n)_{n\geq 1}$ converges in probability to X.*

Proof. If $\mathbb{P}(X = x) = 1$ for some $x \in M$, take $\varepsilon > 0$ and apply (13.3) to a continuous function $f : M \to [0, 1]$ that vanishes on x and equals 1 outside of the open ball of center x and radius $\varepsilon > 0$. □

14. The characteristic function of a random variable

Let X be a random variable on some probability space. We define a complex-valued function $\phi_X : \mathbb{R} \to \mathbb{C}$ by setting

$$\phi_X(t) = E(e^{itX}) = \int_{\mathbb{R}} e^{itx}\, \mathrm{d}\mathbb{P}_X(x),$$

for all $t \in \mathbb{R}$. This is called the *characteristic function* of the random variable X. Note that $|\phi_X(t)| \leq 1$ for all $t \in \mathbb{R}$. Moreover, it follows easily from the Dominated Convergence Theorem that ϕ_X is continuous.

The characteristic function ϕ_X is essentially the same thing as the Fourier transform of the probability measure $\mathbb{P}_X$ regarded as a tempered distribution on the real line. In fact, the Fourier transform of $\mathbb{P}_X$ is given by

$$\mathbb{R} \ni t \longmapsto \frac{1}{\sqrt{2\pi}}\, \phi_X(-t) \in \mathbb{C}.$$

The notion of characteristic function can be readily generalized to random vectors.

Definition 14.1. Let V be a real finite-dimensional vector space endowed with its Borel σ-algebra and X be a V-valued random vector. The *characteristic function* of X is the complex-valued map $\phi_X : V^* \to \mathbb{C}$ defined on the dual space V^* and given by

$$\phi_X(\alpha) = E(e^{i\alpha(X)}) = \int_V e^{i\alpha(x)} \, \mathrm{d}\mathbb{P}_X(x),$$

for all $\alpha \in V^*$.

As before, $|\phi_X(\alpha)| \leq 1$ for all $\alpha \in V^*$ and ϕ_X is continuous. Moreover, up to a sign in α and a multiplicative constant, ϕ_X is simply the Fourier transform of the probability measure $\mathbb{P}_X$ regarded as a tempered distribution.

Since the Fourier transform is injective on tempered distributions and the inclusion of finite countably additive measures in the space of tempered distributions is also injective, we obtain the following result.

Proposition 14.2. *Let V be a real finite-dimensional vector space and X and Y be V-valued random vectors. If $\phi_X = \phi_Y$ then $\mathbb{P}_X = \mathbb{P}_Y$. In particular, if $\alpha(X)$ and $\alpha(Y)$ have the same distribution for all $\alpha \in V^*$ then $\mathbb{P}_X = \mathbb{P}_Y$.* □

15. Conditional probability and independence

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $A, B \in \mathcal{A}$ be events. We wish to define the conditional probability $\mathbb{P}(A|B)$ of A given B. Before presenting the formal definition, we give a motivation in terms of the frequentist interpretation of probability.

Imagine that the random experiment modelled by $(\Omega, \mathcal{A}, \mathbb{P})$ is repeated a large number N of times and, for $C \in \mathcal{A}$, denote by N_C the number of times that the event $\omega \in C$ occurs, where $\omega \in \Omega$ denotes the outcome of the experiment. We then have $\mathbb{P}(C) = \lim_{N\to+\infty} \frac{N_C}{N}$. The conditional probability $\mathbb{P}(A|B)$ should be the limit as $N \to +\infty$ of the frequency with which the event $\omega \in A$ happens among those repetitions of the experiment in which the event $\omega \in B$ happens. Clearly, among the N_B repetitions in which $\omega \in B$ happens, we have that the number of repetitions in which $\omega \in A$ happens is equal to $N_{A\cap B}$. Hence:

$$\mathbb{P}(A|B) = \lim_{N\to+\infty} \frac{N_{A\cap B}}{N_B} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We take the latter quotient as the official definition of conditional probability.

Definition 15.1. Given $A, B \in \mathcal{A}$ with $\mathbb{P}(B) > 0$, the *conditional probability* $\mathbb{P}(A|B)$ of A given B is defined by:

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

Note that the map

$$\mathbb{P}(\,\cdot\,|B) : \mathcal{A} \ni A \longmapsto \mathbb{P}(A|B) \in [0,1] \tag{15.1}$$

is a probability measure on $\mathcal{A}$. This is the probability measure obtained from the following recipe: first, change $\mathbb{P}$ so that it vanishes on the complement of B and remains the same on measurable subsets of B. This yields the measure $\mathcal{A} \ni A \longmapsto \mathbb{P}(A \cap B) \in [0,1]$, which is not in general a probability measure. Now multiply such measure by the appropriate constant to make it a probability measure, obtaining (15.1).

If $X : \Omega \to M$ is a random object taking values in a measurable space $(M, \mathcal{B})$, then for any $A \in \mathcal{A}$ and $B \in \mathcal{B}$ we write

$$\mathbb{P}(A|X \in B) = \frac{\mathbb{P}(A \cap [X \in B])}{\mathbb{P}(X \in B)}$$

for the conditional probability of A given the event $[X \in B]$, provided that $\mathbb{P}(X \in B) > 0$. In particular, if $x \in M$ is such that $\{x\} \in \mathcal{B}$ and $\mathbb{P}(X = x) > 0$, we write $\mathbb{P}(A|X = x)$ for the conditional probability of A given $[X = x]$.

The definition of conditional probability leads naturally to a definition of independence of events. We say that two events A and B are independent if the conditional probability $\mathbb{P}(A|B)$ is equal to $\mathbb{P}(A)$, i.e., the probability of A remains unchanged if we learn that B happened. We rewrite this in a way that $\mathbb{P}(B)$ does not appear in the denominator to avoid the assumption that $\mathbb{P}(B) > 0$.

Definition 15.2. We say that two events $A, B \in \mathcal{A}$ are *independent* if:

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

Clearly A and B are independent if either A or B has a probability of zero or one.

The law of total probability, stated below, is a useful method for computing the probability of an event by "breaking into cases", i.e., we compute the probability of A under various values for a random object X and we combine such probabilities into the probability of A.

Proposition 15.3 (law of total probability)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to M$ be a random object taking values in a countable measurable space $(M, \mathcal{B})$ with $\mathcal{B} = \wp(M)$. Given $A \in \mathcal{A}$, we have that:*

$$\mathbb{P}(A) = \sum_{\substack{x \in M \\ \mathbb{P}(X=x)>0}} \mathbb{P}(A|X = x)\mathbb{P}(X = x). \tag{15.2}$$

Moreover, for any subset B of M we have:

$$\mathbb{P}\big(A \cap [X \in B]\big) = \sum_{\substack{x \in B \\ \mathbb{P}(X=x)>0}} \mathbb{P}(A|X=x)\mathbb{P}(X=x). \tag{15.3}$$

Proof. Simply note that A is the disjoint countable union of $A \cap [X = x]$ with x ranging over M and that $A \cap [X \in B]$ is the disjoint countable union of $A \cap [X = x]$ with x ranging over B. □

If the random object X is not discrete it might happen that $\mathbb{P}(X = x) = 0$ for all $x \in M$, so that the conditional probability $\mathbb{P}(A|X = x)$ will never make sense. Yet we want it to make sense and we want a version of the law of total probability to hold when X is not discrete! We will achieve this by replacing the sums in (15.2) and (15.3) with an integral with respect to $\mathbb{P}_X$ and by defining the conditional probability $\mathbb{P}(A|X = x)$ even when $\mathbb{P}(X = x) = 0$ in a way that forces the law of total probability (with an integral) to hold. More explicitly, equality (15.3) should be replaced by:

$$\mathbb{P}\big(A \cap [X \in B]\big) = \int_B \mathbb{P}(A|X=x)\,\mathrm{d}\mathbb{P}_X(x). \tag{15.4}$$

Note that if $X : \Omega \to M$ is a random object taking values in an arbitrary measurable space $(M, \mathcal{B})$ then for any $A \in \mathcal{A}$ the map

$$\mathcal{B} \ni B \longmapsto \mathbb{P}\big(A \cap [X \in B]\big) \in [0, 1] \tag{15.5}$$

is a finite nonnegative countably additive measure. Moreover, the statement that equality (15.4) holds for all $B \in \mathcal{B}$ is equivalent to the statement that the map

$$\Omega \ni x \longmapsto \mathbb{P}(A|X=x) \in [0, +\infty[$$

be a Radon–Nikodym derivative of the finite measure (15.5) with respect to the probability measure $\mathbb{P}_X$. Clearly, (15.5) is absolutely continuous with respect to $\mathbb{P}_X$ and thus the Radon–Nikodym Theorem guarantees that such derivative exists. Moreover, since

$$\mathbb{P}\big(A \cap [X \in B]\big) \leq \mathbb{P}_X(B)$$

for all $B \in \mathcal{B}$, any such Radon–Nikodym derivative will take values in $[0, 1]$ at $\mathbb{P}_X$-almost every point of M. We may then choose a Radon–Nikodym derivative that takes values in $[0, 1]$ at every point of M.

Definition 15.4. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to M$ be a random object taking values in a measurable space $(M, \mathcal{B})$. For any $A \in \mathcal{A}$, a *conditional probability of A given X* is any measurable map

$$M \ni x \longmapsto \mathbb{P}(A|X=x) \in [0, 1] \tag{15.6}$$

such that (15.4) holds for all $B \in \mathcal{B}$.

We have proven the following result.

Proposition 15.5. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to M$ be a random object taking values in a measurable space $(M, \mathcal{B})$. Given an event $A \in \mathcal{A}$, a conditional probability of A given X exists. Moreover, if* (15.6) *is a conditional probability of A given X then another measurable map*

$$M \ni x \longmapsto \mathbb{P}(A|X = x)' \in [0, 1]$$

is also a conditional probability of A given X if and only if it is equal $\mathbb{P}_X$-almost everywhere to (15.6). □

Thus a conditional probability of A given X is not usually unique and for a specific point $x \in M$, the value of $\mathbb{P}(A|X = x)$ typically depends on the choice of a particular conditional probability of A given X. If $\{x\} \in \mathcal{B}$ and $\mathbb{P}(X = x) > 0$, using (15.4) with $B = \{x\}$ we get

$$\mathbb{P}(A|X = x) = \frac{\mathbb{P}\big(A \cap [X = x]\big)}{\mathbb{P}(X = x)},$$

so that $\mathbb{P}(A|X = x)$ is indeed well-defined in a way that is consistent with Definition 15.1. However, if $\{x\} \in \mathcal{B}$ and $\mathbb{P}(X = x) = 0$ then one can change the value of (15.6) at will at the point x and the new map will remain a valid conditional probability of A given X.

Remark 15.6. Within a purely mathematical point of view, it makes no sense to put any restrictions on the measurable space $(M, \mathcal{B})$ in Definition 15.4, as the definition makes sense and some basic results will hold for arbitrary measurable spaces. However, for practical applications, one has to be careful since for weird choices of $(M, \mathcal{B})$ what we call $\mathbb{P}(A|X = x)$ might not mean what our notation and terminology suggests it to mean. To begin with, if it is not true that all singletons $\{x\}$, $x \in M$, belong to $\mathcal{B}$ then the meaning of $\mathbb{P}(A|X = x)$ is often what one would normally call the probability of A conditioned on $[X \in B]$, with B the atom of $\mathcal{B}$ containing x if it exists (see the discussion in Subsection 19.1). Even when all singletons are in $\mathcal{B}$, the conditional probability $\mathbb{P}(A|X = x)$ might behave in an unexpected way. For example, assume that M is uncountable and that $\mathcal{B}$ is the σ-algebra of subsets of M consisting of all countable subsets of M and of all subsets of M whose complement in M is countable. In this case, if $\mathbb{P}(X = x) = 0$ for all $x \in M$, then one easily checks that setting $\mathbb{P}(A|X = x) = \mathbb{P}(A)$ for all $x \in M$ we obtain a conditional probability of A given X. Thus, we might have situations in which the value of X gives relevant information about the event A, and yet the totally unreasonable equality $\mathbb{P}(A|X = x) = \mathbb{P}(A)$ holds. It is the weird choice of σ-algebra $\mathcal{B}$ that is creating the problem here.

So, what are the "safe" cases in which what we defined as $\mathbb{P}(A|X = x)$ means what one expect it to mean at least for $\mathbb{P}_X$-almost every $x \in M$? As we explain below, a separable metric space M endowed with its Borel σ-algebra $\mathcal{B}$ is a "safe" case.

Let M be a metric space and $\mathcal{B}$ be the Borel σ-algebra of M. If μ and ν are nonnegative countably additive measures on $\mathcal{B}$ that are finite on bounded sets and if ν is absolutely continuous with respect to μ, then for μ-almost every $x \in M$ the value at x of a Radon–Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ can be written as a limit of quotients of the form

$$\frac{\nu(B)}{\mu(B)} \tag{15.7}$$

with $B \in \mathcal{B}$ containing x, the diameter of B tending to zero and (x, B) belonging to what is called a *Vitali relation* $\mathcal{V}$ for μ. See [2, Sections 2.8, 2.9] for details. If M is separable, a Vitali relation $\mathcal{V}$ always exists and in fact, in many cases — for instance, if M is a real finite-dimensional normed vector space — one can take $\mathcal{V}$ simply as the set of pairs (x, B) with B a closed ball centered at x. In the case we are interested, ν is (15.5) and $\mu = \mathbb{P}_X$, so that the quotient (15.7) is simply the standard conditional probability of A given the event $[X \in B]$. Hence $\mathbb{P}(A|X = x)$ will be equal for $\mathbb{P}_X$-almost all $x \in M$ to a limit of conditional probabilities and such limit can reasonably be called the conditional probability of A given $[X = x]$.

A difficulty related to the nonuniqueness of the conditional probability of an event given a random object is that if we choose a conditional probability of A given X for every $A \in \mathcal{A}$, it will not in general be true that

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n | X = x\Big) = \sum_{n=1}^{\infty} \mathbb{P}(A_n|X = x), \tag{15.8}$$

for every $x \in M$ and every sequence $(A_n)_{n\geq 1}$ of disjoint elements of $\mathcal{A}$. It is easily checked that equality (15.8) must holt for $\mathbb{P}_X$-almost every $x \in M$, but the set of probability zero in which the equality fails in general depends on the sequence $(A_n)_{n\geq 1}$ and since there are usually uncountably many such sequences, it is not clear that one can make such set of probability zero independent of $(A_n)_{n\geq 1}$.

Though sets of probability zero are not important, it would nevertheless be nice if the map $\mathcal{A} \ni A \mapsto \mathbb{P}(A|X = x) \in [0, 1]$ were truly a probability measure for all $x \in M$. This leads us to the notion of regular conditional probability.

Definition 15.7. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $X : \Omega \to M$ be a random object taking values in a measurable space $(M, \mathcal{B})$ and $Y : \Omega \to M'$ be a random object taking values in a measurable space $(M', \mathcal{B}')$. A *regular conditional probability for Y given X* is a map

$$M \times \mathcal{B}' \ni (x, B) \longmapsto \mathbb{P}(Y \in B|X = x) \in [0, 1] \tag{15.9}$$

having the following properties:

(i) for every $B \in \mathcal{B}'$, the map $M \ni x \mapsto \mathbb{P}(Y \in B|X = x) \in [0, 1]$ is a conditional probability of $[Y \in B]$ given X, i.e., it is a measurable

map such that

$$\mathbb{P}\big([Y \in B] \cap [X \in C]\big) = \int_C \mathbb{P}(Y \in B | X = x)\, \mathrm{d}\mathbb{P}_X(x),$$

for all $C \in \mathcal{B}$;

(ii) for every $x \in M$, the map $\mathcal{B}' \ni B \mapsto \mathbb{P}(Y \in B | X = x) \in [0,1]$ is a probability measure.

If $(M', \mathcal{B}') = (\Omega, \mathcal{A})$ and Y is the identity map of Ω, we write simply $\mathbb{P}(A|X = x)$ instead of $\mathbb{P}(Y \in A | X = x)$ for all $A \in \mathcal{A}$ and all $x \in M$ and we call the map

$$M \times \mathcal{A} \ni (x, A) \longmapsto \mathbb{P}(A|X = x) \in [0,1]$$

a *regular conditional probability given* X.

Thus, a regular conditional probability given X is the same thing as a choice of conditional probability of A given X for all $A \in \mathcal{A}$ in such a way that equality (15.8) holds for all $x \in M$ and every sequence $(A_n)_{n \geq 1}$ of disjoint elements of $\mathcal{A}$ and $\mathbb{P}(\Omega | X = x) = 1$ for all $x \in M$.

Note that a regular conditional probability (15.9) of Y given X can be identified with a map

$$M \ni x \longmapsto \mathbb{P}(Y \in \cdot\, | X = x) \in \mathrm{Prob}(M', \mathcal{B}') \tag{15.10}$$

taking values in the set $\mathrm{Prob}(M', \mathcal{B}')$ of probability measures on the measurable space $(M', \mathcal{B}')$, where $\mathbb{P}(Y \in \cdot\, | X = x)$ denotes the probability measure $\mathcal{B}' \ni B \longmapsto \mathbb{P}(Y \in B | X = x) \in [0,1]$.

Unfortunately, a regular conditional probability for Y given X does not always exist, though in Section 17 we will show that it does exist under mild assumptions on the measurable space $(M', \mathcal{B}')$ in which Y takes values. That is the reason why we chose to define regular conditional probabilities of Y given X instead of simply defining regular conditional probabilities given X: it could happen that $(M', \mathcal{B}')$ satisfies the assumption that ensures existence of a regular conditional probability of Y given X, while $(\Omega, \mathcal{A})$ does not satisfy such assumption and we cannot be sure that a regular conditional probability given X exists. When a regular conditional probability given X does exist, one can obtain a regular conditional probability of Y given X simply by applying the conditional probability given X to events of the form $[Y \in B]$.

15.1. **Independence of two random objects.** We will say that two random objects X and Y are independent if a regular conditional probability (15.9) of Y given X can be found that is independent of $x \in M$. This is easily seen to be equivalent to the definition below.

Definition 15.8. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $X : \Omega \to M$ be a random object taking values in a measurable space $(M, \mathcal{B})$ and $Y : \Omega \to M'$

be a random object taking values in a measurable space $(M', \mathcal{B}')$. We say that X and Y are *independent* if

$$\mathbb{P}\big([X \in B] \cap [Y \in B']\big) = \mathbb{P}(X \in B)\mathbb{P}(Y \in B'), \tag{15.11}$$

for all $B \in \mathcal{B}$ and all $B' \in \mathcal{B}'$.

In other words, X and Y are independent if and only if every event in the σ-algebra $\big\{X^{-1}[B] : B \in \mathcal{B}\big\}$ induced by X is independent of every event in the σ-algebra $\big\{Y^{-1}[B'] : B' \in \mathcal{B}'\big\}$ induced by Y. Moreover, X and Y are independent if and only if

$$\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y,$$

where $\mathbb{P}_X \otimes \mathbb{P}_Y$ denotes the product of the measures $\mathbb{P}_X$ and $\mathbb{P}_Y$. Using this observation and Lemma 5.1 we conclude that in order to check that X and Y are independent it is sufficient to check that equality (15.11) holds for all B and B' belonging to fixed sets of generators of the σ-algebras $\mathcal{B}$ and $\mathcal{B}'$ that are closed under finite intersections.

The following is a direct consequence of Definition 15.8.

Proposition 15.9. *Let X be a random object taking values in a measurable space $(M, \mathcal{B})$, Y be a random object on the same probability space taking values in a measurable space $(M', \mathcal{B}')$ and let $f : M \to M_1$ and $g : M' \to M_1'$ be measurable maps, where $(M_1, \mathcal{B}_1)$ and $(M_1', \mathcal{B}_1')$ are measurable spaces. If the random objects X and Y are independent then so are the random objects $f(X)$ and $g(Y)$.* □

Using Fubini–Tonelli's Theorem for the product measure $\mathbb{P}_X \otimes \mathbb{P}_Y$ we show the following useful property of the expected value of the product of two independent random variables.

Proposition 15.10. *Let X and Y be independent random variables on the same probability space. If either X and Y are both nonnegative or both X and Y have finite expected value then the expected value of the product XY exists and it is equal to the product of the expected values, i.e.*

$$E(XY) = E(X)E(Y),$$

where the convention $0 \cdot (+\infty) = (+\infty) \cdot 0 = 0$ is used.

Proof. Simply note that

$$E(XY) = \int_{\mathbb{R}^2} xy \, \mathrm{d}\mathbb{P}_{(X,Y)}(x, y) = \int_{\mathbb{R}^2} xy \, \mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_Y)(x, y)$$

and apply Fubini–Tonneli's Theorem. □

Corollary 15.11. *If X and Y are independent square integrable random variables on the same probability space then $\mathrm{Cov}(X, Y) = 0$. Moreover, if X and Y are independent square integrable random vectors on the same probability space taking values in real finite-dimensional vector spaces V and W then $\mathrm{Cov}(X, Y) \in V \otimes W$ is zero.*

Proof. The first part of the statement follows directly from Proposition 15.10 and the second follows from the first by noting that, by Proposition 15.9, the random variables $\alpha(X)$ and $\beta(Y)$ are independent for all $\alpha \in V^*$ and $\beta \in W^*$. □

Note that the proof of Corollary 15.11 does not use the fact that X and Y are square integrable, but we include that assumption in the statement as covariance was only defined for square integrable random variables (or random vectors).

15.2. **The uniqueness problem for regular conditional probability.** Let X be a random object on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ taking values in a measurable space $(M, \mathcal{B})$ and Y be a random object on the same probability space taking values in a measurable space $(M', \mathcal{B}')$. We have seen that a regular conditional probability of Y given X is typically not unique if it exists. Moreover, if

$$M \times \mathcal{B}' \ni (x, B) \longmapsto \mathbb{P}(Y \in B | X = x) \in [0, 1], \tag{15.12}$$

$$M \times \mathcal{B}' \ni (x, B) \longmapsto \mathbb{P}(Y \in B | X = x)' \in [0, 1] \tag{15.13}$$

are both regular conditional probabilities of Y given X then for all $B \in \mathcal{B}'$ the equality

$$\mathbb{P}(Y \in B | X = x) = \mathbb{P}(Y \in B | X = x)' \tag{15.14}$$

holds for $\mathbb{P}_X$-almost every $x \in M$.

As discussed above, the maps (15.12) and (15.13) can be identified with the maps:

$$M \ni x \longmapsto \mathbb{P}(Y \in \cdot\, | X = x) \in \mathrm{Prob}(M', \mathcal{B}'), \tag{15.15}$$

$$M \ni x \longmapsto \mathbb{P}(Y \in \cdot\, | X = x)' \in \mathrm{Prob}(M', \mathcal{B}'). \tag{15.16}$$

Is it true that (15.15) and (15.16) are equal for $\mathbb{P}_X$-almost every $x \in M$? That is equivalent to saying that the set of probability zero in which equality (15.14) fails can be chosen independently of $B \in \mathcal{B}'$. This can indeed be done if $\mathcal{B}'$ is countably generated.

Proposition 15.12. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $X : \Omega \to M$ be a random object taking values in a measurable space $(M, \mathcal{B})$ and $Y : \Omega \to M'$ be a random object taking values in a measurable space $(M', \mathcal{B}')$. If the σ-algebra $\mathcal{B}'$ admits a countable set of generators (for instance, if $\mathcal{B}'$ is the Borel σ-algebra of a second countable topology) then two regular conditional probabilities* (15.15) *and* (15.16) *of Y given X are equal $\mathbb{P}_X$-almost surely.*

Proof. Simply note that a countable set of generators for $\mathcal{B}'$ can be made closed under finite intersections and use Lemma 5.1. □

There are some situations in which one really needs to give a meaning to the conditional probability $\mathbb{P}(Y \in \cdot\, | X = x)$ for a specific point $x \in M$ with $\mathbb{P}(X = x) = 0$, notably in the context of Bayesian statistics. This can be

achieved in many situations by adding suitable continuity requirements to the regular conditional probabilities.

Proposition 15.13. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to M$ and $Y : \Omega \to M'$ be random objects taking values in topological spaces M and M' endowed with their respective Borel σ-algebras $\mathcal{B}$ and $\mathcal{B}'$. Assume that M' is metrizable and separable and let $\mathrm{Prob}(M', \mathcal{B}')$ be endowed with the weak topology. If the regular conditional probabilities* (15.15) *and* (15.16) *of Y given X are continuous at a point $x \in M$ in the support of the measure $\mathbb{P}_X$ then such regular conditional probabilities coincide at the point x.*

Proof. It follows from Proposition 15.12 and from the fact that the weak topology is Hausdorff (Remark 13.2). □

16. Markov kernels and generalized product measures

A regular conditional probability is an example of what we call a Markov kernel.

Definition 16.1. Let $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ be measurable spaces. A *Markov kernel* (or simply *kernel*) with source $(M, \mathcal{B})$ and target $(M', \mathcal{B}')$ is a map

$$K : M \longrightarrow \mathrm{Prob}(M', \mathcal{B}')$$

from M to the set of probability measures on $(M', \mathcal{B}')$ such that the map

$$M \ni x \longmapsto K(x)(B) \in [0, 1] \tag{16.1}$$

is measurable, for every $B \in \mathcal{B}'$.

It follows from Lemma 5.2 that $K : M \to \mathrm{Prob}(M', \mathcal{B}')$ is a Markov kernel if and only if the map (16.1) is measurable for every B in a certain fixed collection of generators of the σ-algebra $\mathcal{B}'$ that is closed under finite intersections. Namely, note that the collection of sets $B \in \mathcal{B}'$ for which the map (16.1) is measurable is a σ-additive class of which M' is a member.

Given a kernel K as in Definition 16.1 and a probability measure defined on $\mathcal{B}$, we can construct a probability measure defined on the product σ-algebra $\mathcal{B} \otimes \mathcal{B}'$ which generalizes the standard construction of a product measure by allowing the measure on the second factor to be a function of the point on the first factor. We need a preparatory lemma.

Lemma 16.2. *Let $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ be measurable spaces and K be a kernel with source $(M, \mathcal{B})$ and target $(M', \mathcal{B}')$. For every element C of the product σ-algebra $\mathcal{B} \otimes \mathcal{B}'$, the map*

$$M \ni x \longmapsto K(x)(C_x) \in [0, 1]$$

is measurable, where:

$$C_x = \big\{ y \in M' : (x, y) \in C \big\}. \tag{16.2}$$

Proof. Follows from Lemma 5.2 by noting that the collection of all $C \in \mathcal{B} \otimes \mathcal{B}'$ for which the thesis holds is a σ-additive class that contains all products $B \times B'$ with $B \in \mathcal{B}$ and $B' \in \mathcal{B}'$. □

Proposition 16.3. *Let $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ be measurable spaces, $\mathbb{P}$ be a probability measure defined on $\mathcal{B}$ and K be a kernel with source $(M, \mathcal{B})$ and target $(M', \mathcal{B}')$. There exists a unique probability measure $\mathbb{P} \star K$ defined on the product σ-algebra $\mathcal{B} \otimes \mathcal{B}'$ such that:*

$$(\mathbb{P} \star K)(B \times B') = \int_B K(x)(B') \, \mathrm{d}\mathbb{P}(x),$$

for all $B \in \mathcal{B}$ and all $B' \in \mathcal{B}'$. Moreover, for every $C \in \mathcal{B} \otimes \mathcal{B}'$, we have:

$$(\mathbb{P} \star K)(C) = \int_M K(x)(C_x) \, \mathrm{d}\mathbb{P}(x), \tag{16.3}$$

where C_x is defined by (16.2).

Proof. Uniqueness follows from Lemma 5.1 and existence follows by noting that formula (16.3) defines a probability measure with the desired property. Lemma 16.2 ensures that the integral in (16.3) is well-defined. □

Obviously, the measure $\mathbb{P} \star K$ is simply the standard product measure in the trivial case in which the kernel K is constant.

Note that $\mathbb{P}$ is the push-forward of $\mathbb{P} \star K$ under the first projection. If X is an $(M, \mathcal{B})$-valued random object and Y is an $(M', \mathcal{B}')$-valued random object on the same probability space as X then it follows directly from the corresponding definitions that a regular conditional probability of Y given X, when identified with the map (15.10), is the same thing as a kernel K with source $(M, \mathcal{B})$ and target $(M', \mathcal{B}')$ such that:

$$\mathbb{P}_{(X,Y)} = \mathbb{P}_X \star K.$$

We can think about the probability measure $\mathbb{P} \star K$ as modelling the process of randomly choosing $(x, y) \in M \times M'$ by first choosing $x \in M$ according to the probability measure $\mathbb{P}$ and then choosing $y \in M'$ according to the probability measure $K(x)$.

The following is a simple consequence of (16.3) and the observation above.

Proposition 16.4. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $X : \Omega \to M$ be a random object taking values in a measurable space $(M, \mathcal{B})$ and $Y : \Omega \to M'$ be a random object taking values in a measurable space $(M', \mathcal{B}')$. Given a regular conditional probability*

$$M \times \mathcal{B}' \ni (x, B) \longmapsto \mathbb{P}(Y \in B | X = x) \in [0, 1]$$

of Y given X we have that

$$\mathbb{P}\big((X, Y) \in C\big) = \int_M \mathbb{P}(Y \in C_x | X = x) \, \mathrm{d}\mathbb{P}_X(x),$$

for any $C \in \mathcal{B} \otimes \mathcal{B}'$, where C_x is defined by (16.2). □

Example 16.5. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to M$ be a random object with $(M, \mathcal{B})$ a measurable space. Assume that all singletons $\{x\}$, $x \in M$, belong to the σ-algebra $\mathcal{B}$. Given a regular conditional

probability of X given X, it seems reasonable that one should have

$$\mathbb{P}(X = x|X = x) = 1, \tag{16.4}$$

for all $x \in M$. This is not in general true as one can replace the probability measure $\mathbb{P}(X \in \cdot\,|X = x)$ by any other probability measure for x in a subset of M with $\mathbb{P}_X$-probability zero (as long as the appropriate measurability requirements are kept). Can we at least show that (16.4) holds for $\mathbb{P}_X$-almost every $x \in M$? In general no (see Example 16.6 below), but the answer is yes under a mild assumption. Assume that the diagonal

$$\Delta_M = \big\{(x, x) : x \in M\big\}$$

belongs to $\mathcal{B} \otimes \mathcal{B}$. This is true, for instance, if $\mathcal{B}$ is the Borel σ-algebra of a Hausdorff second countable topology on M. Proposition 16.4 yields:

$$1 = \mathbb{P}\big((X, X) \in \Delta_M\big) = \int_M \mathbb{P}(X = x|X = x)\,\mathrm{d}\mathbb{P}_X(x)$$

and since $\mathbb{P}(X = x|X = x) \leq 1$ for all $x \in M$ it follows that (16.4) holds for $\mathbb{P}_X$-almost every $x \in M$.

Example 16.6. Let Ω be an uncountable set, $\mathcal{A}$ be the σ-algebra consisting of all countable subsets of Ω and all subsets of Ω with a countable complement in Ω. Define a probability measure $\mathbb{P} : \mathcal{A} \to [0, 1]$ by setting $\mathbb{P}(A) = 0$ if A is countable and $\mathbb{P}(A) = 1$ if the complement of A in Ω is countable. Since $\mathbb{P}$ only takes the values 0 and 1, every pair of events is independent. If $(M, \mathcal{B}) = (\Omega, \mathcal{A})$ and $X : \Omega \to M$ is the identity map, we then have that X is independent of itself. Thus

$$M \times \mathcal{B} \ni (x, B) \longmapsto \mathbb{P}(X \in B|X = x) = \mathbb{P}(X \in B) \in [0, 1]$$

is a regular conditional probability of X given X. Note that

$$\mathbb{P}(X = x|X = x) = \mathbb{P}(X = x) = 0,$$

for all $x \in M$.

There is a Fubini–Tonelli Theorem for the generalized product measures $\mathbb{P} \star K$.

Theorem 16.7. *Let $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ be measurable spaces, $\mathbb{P}$ be a probability measure defined on $\mathcal{B}$ and K be a kernel with source $(M, \mathcal{B})$ and target $(M', \mathcal{B}')$. If $f : M \times M' \to [-\infty, +\infty]$ is a $\mathcal{B} \otimes \mathcal{B}'$-measurable function whose integral with respect to $\mathbb{P} \star K$ exists (i.e., either its positive part or its negative part has a finite integral) then for $\mathbb{P}$-almost every $x \in M$ the integral $\int_{M'} f(x, y)\,\mathrm{d}K(x)(y)$ of $f(x, \cdot)$ with respect to the probability measure $K(x)$*

exists, the function[1]

$$M \ni x \longmapsto \int_{M'} f(x,y)\,\mathrm{d}K(x)(y) \in [-\infty,+\infty] \tag{16.5}$$

is measurable, its integral exists and the equality

$$\int_{M\times M'} f\,\mathrm{d}(\mathbb{P}\star K) = \int_M \left(\int_{M'} f(x,y)\,\mathrm{d}K(x)(y)\right)\mathrm{d}\mathbb{P}(x)$$

holds.

Proof. Follow the standard recipe: prove the result first for indicator functions of measurable subsets using (16.3), then for nonnegative simple measurable functions using linear combinations, then for nonnegative measurable functions f using the Monotone Convergence Theorem and the fact that f is an increasing pointwise limit of nonnegative simple measurable functions. Finally prove the general case by writing f as the difference of its positive and negative part. □

The generalized product can be easily iterated a finite number of times. Let $(M_i,\mathcal{B}_i)$, $i=1,\dots,n$ be measurable spaces, $\mathbb{P}$ be a probability measure defined on $\mathcal{B}_1$ and for each $i=1,\dots,n-1$ let K_i be a kernel with source $\left(\prod_{j=1}^i M_j, \bigotimes_{j=1}^i \mathcal{B}_j\right)$ and target $(M_{i+1},\mathcal{B}_{i+1})$. We define a probability measure $\mathbb{P}\star(K_1,\dots,K_{n-1})$ on $\left(\prod_{j=1}^n M_j, \bigotimes_{j=1}^n \mathcal{B}_j\right)$ recursively by setting:

$$\mathbb{P}\star(K_1,\dots,K_j) = \big(\mathbb{P}\star(K_1,\dots,K_{j-1})\big)\star K_j,$$

for all $j=1,\dots,n-1$, where $\mathbb{P}\star(\,)$ (the product constructed with an empty sequence of kernels) is simply the same as $\mathbb{P}$. We can think about the probability measure $\mathbb{P}\star(K_1,\dots,K_{n-1})$ as modelling the process of randomly choosing a sequence $(x_1,\dots,x_n)\in\prod_{i=1}^n M_i$ by first choosing $x_1\in M_1$ according to the probability measure $\mathbb{P}$, then choosing $x_2\in M_2$ according to the probability measure $K_1(x_1)$, then choosing $x_3\in M_3$ according to the probability measure $K_2(x_1,x_2)$ and so on.

Measure theory textbooks usually construct only products of measure spaces with a finite number of factors but in the case of probability spaces it is possible to construct also infinite products. Moreover, infinite products are actually a necessary foundation for various limit theorems concerning infinite sequences of random variables.

Proposition 16.8. *Let $\big((M_n,\mathcal{B}_n)\big)_{n\geq 1}$ be a sequence of measurable spaces, $\mathbb{P}$ be a probability measure on $(M_1,\mathcal{B}_1)$ and for each $n\geq 1$ let K_n be a kernel with source $\left(\prod_{i=1}^n M_i, \bigotimes_{i=1}^n \mathcal{B}_i\right)$ and target $(M_{n+1},\mathcal{B}_{n+1})$. For each $n\geq 1$, set $\mathbb{P}_n = \mathbb{P}\star(K_1,\dots,K_{n-1})$ (so that $\mathbb{P}_1=\mathbb{P}$). There exists a*

[1]More precisely, once should assign some value to the integral $\int_{M'} f(x,y)\,\mathrm{d}K(x)(y)$ for x in the probability zero set in which such integral does not exist. As the measure is not assumed to be complete, in order to ensure measurability of (16.5), such assignment cannot be too crazy, i.e., one has to choose a measurable function on that set of probability zero for the assignment (for instance, a constant function).

unique probability measure $\mathbb{P}_\infty = \mathbb{P} \star (K_1, K_2, \ldots)$ *on the infinite product* $\big(\prod_{n=1}^\infty M_n, \bigotimes_{n=1}^\infty \mathcal{B}_n\big)$ *such that for all* $n \geq 1$ *the push-forward of* $\mathbb{P}_\infty$ *under the projection*

$$\pi_n : \prod_{i=1}^\infty M_i \ni (x_1, x_2, \ldots) \longmapsto (x_1, x_2, \ldots, x_n) \in \prod_{i=1}^n M_i$$

onto the first n *coordinates is equal to* $\mathbb{P}_n$*.*

Proof. For each $n \geq 1$, denote by $\mathcal{A}_n$ the σ-algebra of subsets of the infinite product $M = \prod_{i=1}^\infty M_i$ induced by π_n. Clearly, $(\mathcal{A}_n)_{n\geq 1}$ is an increasing sequence of σ-algebras of subsets of M and therefore the union $\mathcal{A}_\infty = \bigcup_{n=1}^\infty \mathcal{A}_n$ is an algebra of subsets of M (i.e., it is a nonempty collection of subsets of M closed under finite unions and complements). Moreover, $\mathcal{A}_\infty$ generates the product σ-algebra $\bigotimes_{i=1}^\infty \mathcal{B}_i$.

The requirement that the push-forward of $\mathbb{P}_\infty$ under π_n is equal to $\mathbb{P}_n$ is equivalent to the statement that $\mathbb{P}_\infty$ extends the map $\mathbb{P}^n : \mathcal{A}_n \to [0,1]$ defined by

$$\mathbb{P}^n\big(\pi_n^{-1}[B]\big) = \mathbb{P}_n(B), \tag{16.6}$$

for all $B \in \bigotimes_{i=1}^n \mathcal{B}_i$. Note that the surjectivity of π_n implies that every element of $\mathcal{A}_n$ is of the form $\pi_n^{-1}[B]$ for a unique $B \in \bigotimes_{i=1}^n \mathcal{B}_i$, so that equality (16.6) indeed defines a map $\mathbb{P}^n$ on $\mathcal{A}_n$. Moreover, it is clear that $\mathbb{P}^n$ is a probability measure. We have to check that the maps $\mathbb{P}^n$, $n \geq 1$, admit a common extension $\mathbb{P}_\infty$ to $\mathcal{A}_\infty$. To this aim, for $m \geq n \geq 1$, denote by

$$\pi_{n,m} : \prod_{i=1}^m M_i \longrightarrow \prod_{i=1}^n M_i$$

the projection onto the first n coordinates. Using that $\mathbb{P}_n$ is the push-forward of $\mathbb{P}_{n+1}$ under $\pi_{n,n+1}$ and that $\pi_{n,n+1} \circ \pi_{n+1} = \pi_n$, one obtains that $\mathbb{P}^{n+1}$ extends $\mathbb{P}^n$ for all $n \geq 1$ and this implies that the common extension $\mathbb{P}_\infty : \mathcal{A}_\infty \to [0,1]$ of all $\mathbb{P}^n$ exists. Moreover, $\mathbb{P}_\infty$ is finitely additive.

We will prove that the finitely additive measure $\mathbb{P}_\infty$ is actually countably additive and then Carathéodory's Extension Theorem yields a countably additive extension of $\mathbb{P}_\infty$ to the σ-algebra generated by $\mathcal{A}_\infty$. Such extension is the probability measure $\mathbb{P}_\infty$ whose existence is asserted by the statement of the proposition and the proof will be concluded. To establish the countable additivity of $\mathbb{P}_\infty$, it is sufficient to show that $\lim_{k\to+\infty} \mathbb{P}_\infty(B^k) = 0$ for any decreasing sequence of sets $(B^k)_{k=1}^\infty$ in $\mathcal{A}_\infty$ with $\bigcap_{k=1}^\infty B^k = \emptyset$. We start by introducing some notation to make the exposition cleaner.

For each $n \geq 1$ we denote by $\mathcal{M}_n$ the set of all $[0,1]$-valued measurable maps on $\big(\prod_{i=1}^n M_i, \bigotimes_{i=1}^n \mathcal{B}_i\big)$, by $\mathcal{M}$ the set of all $[0,1]$-valued measurable maps on $\big(M, \bigotimes_{i=1}^\infty \mathcal{B}_i\big)$ and by

$$\pi_{n,m}^* : \mathcal{M}_n \longrightarrow \mathcal{M}_m, \quad \pi_n^* : \mathcal{M}_n \longrightarrow \mathcal{M}$$

the right-composition maps defined by $\pi^*_{n,m}(f) = f \circ \pi_{n,m}$ and $\pi^*_n(f) = f \circ \pi_n$, respectively, for all $m \geq n \geq 1$ and all $f \in \mathcal{M}_n$. We also denote by $\mathcal{M}_\infty$ the union of all the images of the injective maps π^*_n, $n \geq 1$. Note that $\mathcal{M}_\infty$ contains the indicator functions of the sets that belong to $\mathcal{A}_\infty$. Let

$$\rho_{n,n+1} : \mathcal{M}_{n+1} \longrightarrow \mathcal{M}_n$$

denote for each $n \geq 1$ the map given by integration with respect to the $(n+1)$-th variable, i.e.:

$$\rho_{n,n+1}(f)(x) = \int_{M_{n+1}} f(x,y)\, \mathrm{d}K_n(x)(y),$$

for all $f \in \mathcal{M}_{n+1}$ and all $x \in \prod_{i=1}^n M_i$. It follows from the generalized Fubini–Tonelli's Theorem 16.7 that $\rho_{n,n+1}$ is well-defined and preserves integrals, i.e., the integral of f with respect to $\mathbb{P}_{n+1}$ is equal to the integral of $\rho_{n,n+1}(f)$ with respect to $\mathbb{P}_n$. Moreover, the map $\rho_{n,n+1}$ preserves the pointwise partial order of functions and it commutes with pointwise limits by the Dominated Convergence Theorem. The fact that $K_n(x)$ is a probability measure implies that $\rho_{n,n+1}$ is a left inverse for the map $\pi^*_{n,n+1}$.

More generally, for $m \geq n \geq 1$, we define $\rho_{n,m} : \mathcal{M}_m \to \mathcal{M}_n$ by letting $\rho_{n,n}$ be the identity of $\mathcal{M}_n$ and by setting:

$$\rho_{n,m} = \rho_{n,n+1} \circ \rho_{n+1,n+2} \circ \cdots \circ \rho_{m-1,m},$$

so that $\rho_{n,m}$ preserves integrals, preserves the pointwise partial order of functions, commutes with pointwise limits and is a left inverse for $\pi^*_{n,m}$.

For each $n \geq 1$ we now want to define a map

$$\rho_n : \mathcal{M}_\infty \longrightarrow \mathcal{M}_n$$

that is a common extension of all $\rho_{n,m}$ in the sense that

$$\rho_n \circ \pi^*_m = \rho_{n,m}$$

for all $m \geq n$. The existence of ρ_n will follow if we check the compatibility condition

$$\rho_{n,m+1} \circ \pi^*_{m,m+1} = \rho_{n,m}$$

for all $m \geq n$. Such equality is easily obtained as follows:

$$\rho_{n,m+1} \circ \pi^*_{m,m+1} = \rho_{n,m} \circ \rho_{m,m+1} \circ \pi^*_{m,m+1} = \rho_{n,m}.$$

Clearly ρ_n preserves the pointwise partial order of functions. We claim that

$$\rho_{n,n+1} \circ \rho_{n+1} = \rho_n, \tag{16.7}$$

$$\int_{\prod_{i=1}^n M_i} \rho_n(\mathbf{1}_B)\, \mathrm{d}\mathbb{P}_n = \mathbb{P}_\infty(B), \tag{16.8}$$

for all $n \geq 1$ and all $B \in \mathcal{A}_\infty$. To prove (16.7), note that both sides of the equality become equal when composed on the right with π^*_m for any $m \geq n+1$ and to prove (16.8) use that $\mathbf{1}_B = \pi^*_m(\mathbf{1}_C)$ for some $C \in \bigotimes_{i=1}^m \mathcal{B}_i$ and some $m \geq n$.

We have completed the introduction of all the necessary notation and are ready to continue the proof. Let $(B^k)_{k\geq 1}$ be a decreasing sequence of sets in $\mathcal{A}_\infty$ with $\bigcap_{k=1}^\infty B^k = \emptyset$ and assume by contradiction that:

$$\lim_{k\to+\infty} \mathbb{P}_\infty(B^k) > 0.$$

For all $n \geq 1$ and all $k \geq 1$, set $f_n^k = \rho_n(\mathbf{1}_{B^k})$, so that $(f_n^k)_{k\geq 1}$ is a pointwise decreasing sequence in $\mathcal{M}_n$ and thus we can define $f_n \in \mathcal{M}_n$ as the pointwise limit $f_n = \lim_{k\to+\infty} f_n^k$. It follows from (16.7) that

$$\rho_{n,n+1}(f_{n+1}^k) = f_n^k$$

for all $k \geq 1$ and thus taking the pointwise limit as $k \to +\infty$ we obtain

$$\rho_{n,n+1}(f_{n+1}) = f_n,$$

for all $n \geq 1$. Moreover, the Dominated Convergence Theorem and (16.8) yield

$$\int_{\prod_{i=1}^n M_i} f_n \,\mathrm{d}\mathbb{P}_n = \lim_{k\to+\infty} \int_{\prod_{i=1}^n M_i} f_n^k \,\mathrm{d}\mathbb{P}_n = \lim_{k\to+\infty} \mathbb{P}_\infty(B_k) > 0, \tag{16.9}$$

for all $n \geq 1$. We now construct by recursion a sequence $(x_n)_{n\geq 1}$ in $\prod_{n=1}^\infty M_n$ such that

$$f_n(x_1, \ldots, x_n) > 0, \tag{16.10}$$

for all $n \geq 1$. Using (16.9) with $n = 1$ we obtain that $f_1(x_1) > 0$ for some $x_1 \in M_1$. Assuming that $x_1, \ldots, x_n$ have been chosen satisfying (16.10), we have

$$\begin{aligned}\int_{M_{n+1}} f_{n+1}(x_1, \ldots, x_n, y) \,\mathrm{d}K_n(x_1, \ldots, x_n)(y) &= \rho_{n,n+1}(f_{n+1})(x_1, \ldots, x_n) \\ &= f_n(x_1, \ldots, x_n) > 0\end{aligned}$$

and therefore there exists $x_{n+1} \in M_{n+1}$ with $f_{n+1}(x_1, \ldots, x_n, x_{n+1}) > 0$.

To conclude the proof, let us show that $x = (x_n)_{n\geq 1}$ is in B^k for all $k \geq 1$, which will yield a contradiction. Given $k \geq 1$, we have $B_k = \pi_n^{-1}[C]$ for some $C \in \bigotimes_{i=1}^n \mathcal{B}_i$ and some $n \geq 1$, so that $\mathbf{1}_{B_k} = \pi_n^*(\mathbf{1}_C)$, $f_n^k = \rho_n(\mathbf{1}_{B_k}) = \mathbf{1}_C$ and

$$0 < f_n(x_1, \ldots, x_n) \leq f_n^k(x_1, \ldots, x_n) = \mathbf{1}_C(x_1 \ldots, x_n) = \mathbf{1}_{B_k}(x). \qquad \square$$

16.1. **Compositions of kernels with measurable maps.** There are two natural ways of composing a kernel with a measurable map. The first way is composition on the right: if K is a kernel and f is a measurable map taking values in the source of K, then the composition $K \circ f$ is a kernel. The second way is composition on the left with the map that does the push-forward operation: if f is a measurable map defined on the target of the kernel K, then $f_* \circ K$ is a kernel, where $f_* : \mathbb{P} \mapsto f_*\mathbb{P}$ is the map that does the push-forward of probability measures under f.

We have the following simple results.

Proposition 16.9. *Let $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ be measurable spaces, K be a kernel with souce $(M, \mathcal{B})$ and target $(M', \mathcal{B}')$ and $f : M'' \to M$ be a measurable map, where $(M'', \mathcal{B}'')$ is a measurable space. If $\mathbb{P}$ is a probability measure on $\mathcal{B}''$ then*

$$(f_*\mathbb{P}) \star K = (f \times \mathrm{Id})_*[\mathbb{P} \star (K \circ f)],$$

where $f \times \mathrm{Id} : M'' \times M' \to M \times M'$ is defined by $(f \times \mathrm{Id})(x, y) = \big(f(x), y\big)$, for all $x \in M''$ and all $y \in M'$.

Proof. Simply note that both sides of the equality agree on sets of the form $B \times B'$, with $B \in \mathcal{B}$ and $B' \in \mathcal{B}'$. □

Proposition 16.10. *Let $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ be measurable spaces, K be a kernel with souce $(M, \mathcal{B})$ and target $(M', \mathcal{B}')$ and $f : M' \to M''$ be a measurable map, where $(M'', \mathcal{B}'')$ is a measurable space. If $\mathbb{P}$ is a probability measure on $\mathcal{B}$ then*

$$(\mathrm{Id} \times f)_*(\mathbb{P} \star K) = \mathbb{P} \star (f_* \circ K),$$

where $\mathrm{Id} \times f : M \times M' \to M \times M''$ is defined by $(\mathrm{Id} \times f)(x, y) = \big(x, f(y)\big)$, for all $x \in M$ and all $y \in M'$.

Proof. Simply note that both sides of the equality agree on sets of the form $B \times B''$, with $B \in \mathcal{B}$ and $B'' \in \mathcal{B}''$. □

Corollary 16.11. *Let $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ be measurable spaces, X be a random object on a probability space taking values in $(M, \mathcal{B})$ and Y be a random object on that same probability space taking values in $(M', \mathcal{B}')$. If $f : M' \to M''$ is a measurable map, where $(M'', \mathcal{B}'')$ is a measurable space, and K is a regular conditional probability of Y given X then $f_* \circ K$ is a regular conditional probability of $f(Y)$ given X.*

Proof. If K is a conditional probability of Y given X then $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \star K$ and therefore:

$$\mathbb{P}_{(X,f(Y))} = (\mathrm{Id} \times f)_* \, \mathbb{P}_{(X,Y)} = (\mathrm{Id} \times f)_*(\mathbb{P}_X \star K) = \mathbb{P}_X \star (f_* \circ K),$$

which yields the conclusion. □

16.2. **Iterated conditioning.** The star operation $\mathbb{P} \star K$ that creates a generalized product measure using a probability measure $\mathbb{P}$ and a kernel K satisfies an associative property which is easy to formulate and prove. Such associative property has an important interpretation in terms of regular conditional probabilities which roughly states that conditioning a random object Z first on $X = x$ and then on $Y = y$ is the same as conditioning Z on $(X, Y) = (x, y)$. We give the details below.

In order to formulate the associative property for the star operation we need a suitable notion of the star product for two kernels.

Definition 16.12. Let $(M,\mathcal{B})$, $(M',\mathcal{B}')$ and $(M'',\mathcal{B}'')$ be measurable spaces, K be a kernel with source $(M,\mathcal{B})$ and target $(M',\mathcal{B}')$ and K' be a kernel with source $(M\times M',\mathcal{B}\otimes\mathcal{B}')$ and target $(M'',\mathcal{B}'')$. We define $K\star K'$ as the kernel with source $(M,\mathcal{B})$ and target $(M'\times M'',\mathcal{B}'\otimes\mathcal{B}'')$ given by

$$(K\star K')(x)=K(x)\star K'(x,\cdot)\in\mathrm{Prob}(M'\times M'',\mathcal{B}'\otimes\mathcal{B}''),$$

for all $x\in M$, where $K'(x,\cdot)$ denotes the kernel with source $(M',\mathcal{B}')$ and target $(M'',\mathcal{B}'')$ given by $M'\ni y\mapsto K'(x,y)\in\mathrm{Prob}(M'',\mathcal{B}'')$.

To see that $K\star K'$ is indeed a kernel, check that the function

$$M\ni x\longmapsto (K\star K')(x)(B_1\times B_2)=\int_{B_1}K'(x,y)(B_2)\,\mathrm{d}K(x)(y)\in[0,1]$$

is measurable for all $B_1\in\mathcal{B}'$ and all $B_2\in\mathcal{B}''$ by using Theorem 16.7 with $f(x,y)=K'(x,y)(B_2)\,\mathbf{1}_{B_1}(y)$.

Proposition 16.13. *Let $(M,\mathcal{B})$, $(M',\mathcal{B}')$ be measurable spaces, K be a kernel with source $(M,\mathcal{B})$ and target $(M',\mathcal{B}')$ and K' be a kernel with source $(M\times M',\mathcal{B}\otimes\mathcal{B}')$ and target $(M'',\mathcal{B}'')$, where $(M'',\mathcal{B}'')$ is a measurable space. For any probability measure $\mathbb{P}$ on $(M,\mathcal{B})$, we have:*

$$(\mathbb{P}\star K)\star K'=\mathbb{P}\star(K\star K'). \tag{16.11}$$

Proof. Both sides of the equality (16.11) are probability measures on the product $\mathcal{B}\otimes\mathcal{B}'\otimes\mathcal{B}''$ and thus, by Proposition 6.6, it is sufficient to check that they coincide on $B\times B'\times B''$, for all $B\in\mathcal{B}$, $B'\in\mathcal{B}'$ and $B''\in\mathcal{B}''$. This follows by applying Theorem 16.7 with $f(x,y)=K'(x,y)(B'')\mathbf{1}_{B'}(y)\mathbf{1}_B(x)$. □

Corollary 16.14 (conditional law of total probability)**.** *Let X, Y and Z be random objects on the same probability space, with X taking values in a measurable space $(M,\mathcal{B})$, Y taking values in a measurable space $(M',\mathcal{B}')$ and Z taking values in a measurable space $(M'',\mathcal{B}'')$. Let*

$$K(x)(B')=\mathbb{P}(Y\in B'|X=x),\quad x\in M,\ B'\in\mathcal{B}'$$

be a regular conditional probability of Y given X and

$$K'(x,y)(B'')=\mathbb{P}\big(Z\in B''|(X,Y)=(x,y)\big),\quad x\in M,\ y\in M',\ B''\in\mathcal{B}''$$

be a regular conditional probability of Z given (X,Y). We have that $K\star K'$ is a regular conditional probability of (Y,Z) given X; using the notation

$$\mathbb{P}\big((Y,Z)\in C|X=x\big)=(K\star K')(x)(C),\quad x\in M,\ C\in\mathcal{B}'\otimes\mathcal{B}''$$

we have the equality:

$$\mathbb{P}\big((Y,Z)\in B'\times B''|X=x\big)=\int_{B'}\mathbb{P}\big(Z\in B''|(X,Y)=(x,y)\big)\,\mathrm{d}K(x)(y), \tag{16.12}$$

for all $x \in M$, $B' \in \mathcal{B}'$ and $B'' \in \mathcal{B}''$. Moreover, the equality

$$\mathbb{P}(Z \in B''|X = x) = \int_{M'} \mathbb{P}\big(Z \in B''|(X,Y) = (x,y)\big)\, \mathrm{d}K(x)(y), \quad x \in M,\ B'' \in \mathcal{B}'' \tag{16.13}$$

defines a regular conditional probability of Z given X.

Proof. We have

$$\mathbb{P}_X \star (K \star K') = (\mathbb{P}_X \star K) \star K' = \mathbb{P}_{(X,Y)} \star K' = \mathbb{P}_{(X,Y,Z)},$$

which implies that $K \star K'$ is a regular conditional probability of (Y,Z) given X. Equality (16.12) follows by noting that the righthand side of such equality is simply the definition of $(K \star K')(x)(B' \times B'')$. From the fact that $K \star K'$ is a regular conditional probability of (Y,Z) given X, we obtain that $(\pi_2)_* \circ (K \star K')$ is a regular conditional probability of Z given X, where π_2 denotes the second projection of the product $M' \times M''$ (Corollary 16.11). Finally, the righthand side of equality (16.13) is simply $\big((\pi_2)_* \circ (K \star K')\big)(x)(B'')$ and this concludes the proof. □

Recall that the original law of total probability — which became simply the definition of conditional probability given the value of a random object — says that $\mathbb{P}\big((Y,Z) \in B' \times B''\big) = \mathbb{P}\big([Y \in B'] \cap [Z \in B'']\big)$ is obtained by integrating the conditional probability $\mathbb{P}(Z \in B''|Y = y)$ over $y \in B'$ with respect to the distribution of Y. Equality (16.12), which we call the conditional law of total probability, is the same thing but with everything conditioned on $X = x$.

Now let us see how Corollary 16.14, which is a consequence of the associative property (16.11), can be interpreted in terms of iterated conditioning of a random object. Let X, Y, Z, K and K' be as in the statement of Corollary 16.14, so that $K \star K'$ is a regular conditional probability of (Y,Z) given X. For each $x \in M$, consider a pair of random objects

$$Y_{|X=x}, \quad Z_{|X=x} \tag{16.14}$$

whose joint distribution is $(K \star K')(x)$. Concretely, one can simply take (16.14) as the projections of $M' \times M''$, with $\mathcal{B}' \otimes \mathcal{B}''$ endowed with the probability measure $(K \star K')(x)$. We then have that the joint distribution of (16.14) can be understood as being obtained from the distribution of (Y,Z) by conditioning on $X = x$.

Since $(K \star K')(x) = K(x) \star K'(x,\cdot)$, we have that $K'(x,\cdot)$ is a regular conditional probability of $Z_{|X=x}$ given $Y_{|X=x}$. This means that, for $y \in M'$, the probability measure $K'(x,y)$ on $\mathcal{B}''$ can be understood as obtained by conditioning Z first on $X = x$ and then on $Y = y$. However, K' is a regular conditional probability of Z given (X,Y), so that $K'(x,y)$ can also be understood as obtained by conditioning Z on $(X,Y) = (x,y)$.

17. Standard Borel spaces

Despite the fact that the general theory of probability spaces and random objects can be developed to some extent for completely arbitrary measurable spaces, we need to focus on a smaller class of measurable spaces to avoid certain problems after establishing some basic fundamental facts. A useful class of measurable spaces to work with is the class of standard Borel spaces, as it is both well-behaved and sufficiently general to encompass everything necessary for applications.

Definition 17.1. A *Polish space* is a separable topological space whose topology is induced by some complete metric. A *standard Borel space* is a measurable space that is isomorphic to a Borel subset of a Polish space endowed with its Borel σ-algebra.

Recall that an isomorphism of measurable spaces is a bijective measurable map whose inverse is also measurable.

While it may look like the class of standard Borel spaces is really large, it turns out that modulo isomorphisms the class is really small.

Theorem 17.2. *Every two uncountable standard Borel spaces are isomorphic.*

Proof. See [6, Theorem 3.3.13]. □

We can now prove an existence result for regular conditional probabilities.

Theorem 17.3. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $X : \Omega \to M$ and $Y : \Omega \to M'$ be random objects, where $(M, \mathcal{B})$ and $(M', \mathcal{B}')$ are measurable spaces. If $(M', \mathcal{B}')$ is a standard Borel space then there exists a regular conditional probability of Y given X. In particular, if $(\Omega, \mathcal{A})$ is a standard Borel space then there exists a regular conditional probability given X.*

Proof. If M' is countable, we pick for each $y \in M'$ a conditional probability of the event $[Y = y]$ given X

$$M \ni x \longmapsto \mathbb{P}(Y = y | X = x) \in [0, 1]$$

and we set

$$\mathbb{P}(Y \in B | X = x) = \sum_{y \in B} \mathbb{P}(Y = y | X = x),$$

for all $x \in M$ and all $B \in \mathcal{B}' = \wp(M')$. This defines a regular conditional probability of Y given X, except for the fact that the condition

$$\mathbb{P}(Y \in M' | X = x) = 1$$

might fail for x in a subset of M with $\mathbb{P}_X$-probability zero. This is easily fixed, for instance, by simply replacing $\mathbb{P}(Y \in \cdot \,| X = x)$ with some fixed arbitrary probability measure (say, a Dirac delta) defined on $\mathcal{B}'$ for x in that subset with $\mathbb{P}_X$-probability zero.

Assume now that M' is uncountable, so that by Theorem 17.2 we can assume that $M' = \mathbb{R}$ and $\mathcal{B}'$ is the Borel σ-algebra of $\mathbb{R}$. Since a probability measure on the Borel σ-algebra of $\mathbb{R}$ is determined by its cumulative distribution function, that is all we need to define for every $x \in M$.

For each rational number $y \in \mathbb{Q}$, choose a conditional probability of the event $[Y \leq y]$ given X:

$$M \ni x \longmapsto \mathbb{P}(Y \leq y | X = x) \in [0,1].$$

By adjusting the value of $\mathbb{P}(Y \leq y | X = x)$ for x in a subset of M with $\mathbb{P}_X$-probability zero, we can assume that the map

$$\mathbb{Q} \ni y \longmapsto \mathbb{P}(Y \leq y | X = x) \in [0,1] \tag{17.1}$$

is increasing, right continuous and satisfies

$$\lim_{y \to -\infty} \mathbb{P}(Y \leq y | X = x) = 0, \quad \lim_{y \to +\infty} \mathbb{P}(Y \leq y | X = x) = 1,$$

for all $x \in M$. Namely, note that right-continuity at a point $y \in \mathbb{Q}$ of an increasing function $F : \mathbb{Q} \to \mathbb{R}$ is equivalent to $\lim_{n \to +\infty} F(y_n) = F(y)$ for one *specific* decreasing sequence $(y_n)_{n \geq 1}$ in $]y, +\infty[\cap \mathbb{Q}$ with $\lim_{n \to +\infty} y_n = y$ and thus what we are demanding of the map (17.1) can be expressed in terms of a countable number of conditions.

Now, every increasing right continuous function $F : \mathbb{Q} \to \mathbb{R}$ satisfying $\lim_{y \to -\infty} F(y) = 0$ and $\lim_{y \to +\infty} F(y) = 1$ has a unique increasing right continuous extension to all of $\mathbb{R}$ and such extension is the cumulative distribution function of a unique probability measure defined on the Borel σ-algebra of $\mathbb{R}$. We thus obtain a kernel K with source $(M, \mathcal{B})$ and target $(M', \mathcal{B}')$ by letting $K(x)$ be the probability measure whose cumulative distribution function extends (17.1), for all $x \in M$. We then have that $\mathbb{P}_X \star K$ agrees with $\mathbb{P}_{(X,Y)}$ on sets of the form $B \times]-\infty, y]$ with $B \in \mathcal{B}$ and $y \in \mathbb{Q}$ and it follows from Lemma 5.1 that $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \star K$. □

18. Independence of arbitrary families of random objects

The notion of independence of a pair of random objects given in Definition 15.8 can be generalized to arbitrary families of random objects. Recall that two random objects X and Y are independent if and only if their joint distribution $\mathbb{P}_{(X,Y)}$ coincides with the product $\mathbb{P}_X \otimes \mathbb{P}_Y$ of the distributions of X and Y. In order to generalize the notion of independence to arbitrary families it is then convenient to first generalize the notion of product of probability measures to arbitrary families. The hard work has already been done in Proposition 16.8 which implies the existence of countable products of probability measures. The existence of products of arbitrary cardinality then follows from the easy fact proven below that a consistent family of probability measures defined in the products of countable subfamilies can be glued together into a probability measure in the product of an arbitrary family.

Lemma 18.1. *Let $\big((M_i, \mathcal{B}_i)\big)_{i\in I}$ be an arbitrary family of measurable spaces and let the cartesian product*

$$M = \prod_{i\in I} M_i$$

be endowed with the product σ-algebra $\mathcal{B} = \bigotimes_{i\in I} \mathcal{B}_i$. For each countable subset E of I, denote by M_E the countable cartesian product $\prod_{i\in E} M_i$ endowed with the σ-algebra $\mathcal{B}_E = \bigotimes_{i\in E} \mathcal{B}_i$ and by π_E the projection:

$$\pi_E : M \ni (x_i)_{i\in I} \longmapsto (x_i)_{i\in E} \in M_E.$$

Assume also that for each countable subset E of I we are given a probability measure $\mathbb{P}_E$ on $\mathcal{B}_E$ satisfying the following consistency condition: for every pair of countable subsets E and E' of I with $E \subset E'$, it holds that

$$(\pi_{E,E'})_* \mathbb{P}_{E'} = \mathbb{P}_E, \tag{18.1}$$

where $\pi_{E,E'}$ denotes the projection:

$$\pi_{E,E'} : M_{E'} \ni (x_i)_{i\in E'} \longmapsto (x_i)_{i\in E} \in M_E.$$

Under these conditions, there exists a unique probability measure $\mathbb{P}$ on $\mathcal{B}$ such that $(\pi_E)_* \mathbb{P} = \mathbb{P}_E$, for every countable subset E of I.

Proof. For each countable subset E of I, let $\mathcal{A}_E$ be the σ-algebra induced by π_E. It is easy to see that the union of all $\mathcal{A}_E$, with E ranging over all countable subsets of I, is a σ-algebra and therefore it is equal to $\mathcal{B}$. For each countable subset E of I, the fact that π_E is surjective implies that every element of $\mathcal{A}_E$ is of the form $\pi_E^{-1}[B]$ for a unique $B \in \mathcal{B}_E$ and therefore we obtain a probability measure $\mathbb{P}^E$ on $\mathcal{A}_E$ by setting

$$\mathbb{P}^E\big(\pi_E^{-1}[B]\big) = \mathbb{P}_E(B),$$

for all $B \in \mathcal{B}_E$. The condition $(\pi_E)_* \mathbb{P} = \mathbb{P}_E$ on the probability measure $\mathbb{P}$ that we want to define is equivalent to the condition that $\mathbb{P}$ extends $\mathbb{P}^E$. Noting that $\pi_E = \pi_{E,E'} \circ \pi_{E'}$ one readily checks that the consistency condition (18.1) means that $\mathbb{P}^{E'}$ extends $\mathbb{P}^E$, for every pair of countable subsets E, E' of I with $E \subset E'$. Hence there exists a unique map $\mathbb{P} : \mathcal{B} \to [0,1]$ that extends $\mathbb{P}^E$ for every countable subset E of I and it is easily seen that $\mathbb{P}$ is a probability measure. □

Proposition 18.2. *Let $\big((M_i, \mathcal{B}_i, \mathbb{P}_i)\big)_{i\in I}$ be an arbitrary family of probability spaces and consider the product $M = \prod_{i\in I} M_i$ endowed with the product σ-algebra $\mathcal{B} = \bigotimes_{i\in I} \mathcal{B}_i$. There exists a unique probability measure $\mathbb{P}$ defined on $\mathcal{B}$ such that*

$$\mathbb{P}\big(\pi_{i_1}^{-1}[B_1] \cap \ldots \cap \pi_{i_n}^{-1}[B_n]\big) = \mathbb{P}_{i_1}(B_1) \cdots \mathbb{P}_{i_n}(B_n), \tag{18.2}$$

for any distinct $i_1, \ldots, i_n \in I$, any $B_1 \in \mathcal{B}_{i_1}$, …, $B_n \in \mathcal{B}_{i_n}$ and any $n \geq 1$, where $\pi_i : M \to M_i$ denotes the projection onto the i-th coordinate for all $i \in I$.

Proof. Uniqueness follows from Proposition 6.5. To prove existence, we will use Lemma 18.1. For every countable subset E of I, Proposition 16.8 yields that there exists a unique probability measure $\mathbb{P}_E$ on $\mathcal{B}_E$ such that

$$\mathbb{P}_E\big(\pi_{E,i_1}^{-1}[B_1] \cap \ldots \cap \pi_{E,i_n}^{-1}[B_n]\big) = \mathbb{P}_{i_1}(B_1) \cdots \mathbb{P}_{i_n}(B_n), \tag{18.3}$$

for any distinct $i_1, \ldots, i_n \in E$, any $B_1 \in \mathcal{B}_{i_1}$, …, $B_n \in \mathcal{B}_{i_n}$ and any $n \geq 1$, where $\pi_{E,i} : M_E \to M_i$ denotes the projection onto the i-th coordinate for all $i \in E$. Namely, simply choose an arbitrary enumeration of E and use the particular case of Proposition 16.8 in which the kernels are constant. The validity of the consistency condition (18.1) follows from the fact that $(\pi_{E,E'})_* \, \mathbb{P}_{E'}$ satisfies the condition (18.3) that characterizes $\mathbb{P}_E$ and then the probability measure $\mathbb{P}$ given by Lemma 18.1 satisfies the condition in the statement of the proposition. □

Definition 18.3. The probability measure $\mathbb{P}$ whose existence and uniqueness is established by Proposition 18.2 is called the *product* of the family of probability measures $(\mathbb{P}_i)_{i \in I}$ and it is denoted by $\bigotimes_{i \in I} \mathbb{P}_i$.

Note that condition (18.2) implies in particular that $\mathbb{P}_i$ is the push-forward of the product $\mathbb{P}$ under the i-th projection π_i. More generally, if $J \subset I$ is any subset, then $\bigotimes_{i \in J} \mathbb{P}_i$ is the push-forward of the product $\mathbb{P} = \bigotimes_{i \in I} \mathbb{P}_i$ under the projection

$$\prod_{i \in I} M_i \ni (x_i)_{i \in I} \longmapsto (x_i)_{i \in J} \in \prod_{i \in J} M_i \tag{18.4}$$

as such push-forward satisfies the condition that characterizes the product $\bigotimes_{i \in J} \mathbb{P}_i$.

Definition 18.4. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(X_i)_{i \in I}$ be a family of random objects $X_i : \Omega \to M_i$, with $(M_i, \mathcal{B}_i)$ a measurable space for all $i \in I$. The family $(X_i)_{i \in I}$ is said to be *independent* if its joint distribution $\mathbb{P}_{(X_i)_{i \in I}}$ coincides with the product $\bigotimes_{i \in I} \mathbb{P}_{X_i}$.

Note that one can equivalently define that $(X_i)_{i \in I}$ is independent if the joint distribution $\mathbb{P}_{(X_i)_{i \in I}}$ is equal to *some* product $\bigotimes_{i \in I} \mathbb{P}_i$ of probability measures $\mathbb{P}_i : \mathcal{B}_i \to [0, 1]$, as taking the push-forward under the i-th projection we obtain that necessarily $\mathbb{P}_i = \mathbb{P}_{X_i}$. Moreover, if $(X_i)_{i \in I}$ is independent then for every $J \subset I$ the subfamily $(X_i)_{i \in J}$ is also independent as the map $(X_i)_{i \in J}$ is the composition of the map $(X_i)_{i \in I}$ with the projection (18.4).

It follows directly from the definition of the product of a family of probability measures that $(X_i)_{i \in I}$ is independent if and only if

$$\mathbb{P}\big([X_{i_1} \in B_1] \cap \ldots \cap [X_{i_n} \in B_n]\big) = \mathbb{P}(X_{i_1} \in B_1) \cdots \mathbb{P}(X_{i_n} \in B_n), \tag{18.5}$$

for any distinct $i_1, \ldots, i_n \in I$, any $B_1 \in \mathcal{B}_{i_1}$, …, $B_n \in \mathcal{B}_{i_n}$ and any $n \geq 1$. In particular, $(X_i)_{i \in I}$ is independent if and only if $(X_i)_{i \in F}$ is independent for any finite subset F of I. Note also that Proposition 6.5 yields that in order to check that $(X_i)_{i \in I}$ is independent it is sufficient to verify equality (18.5) for sets B_1, …, B_n belonging to fixed collections of generators $\mathcal{C}_i$ for

the σ-algebras $\mathcal{B}_i$ that are closed under finite intersections. If I is finite, Proposition 6.6 gives the following more convenient formulation: $(X_i)_{i\in I}$ is independent if and only if

$$\mathbb{P}\Big(\bigcap_{i\in I}[X_i \in B_i]\Big) = \prod_{i\in I}\mathbb{P}(X_i \in B_i),$$

for any choice of $B_i \in \mathcal{C}_i \cup \{M_i\}$, $i \in I$, where $\mathcal{C}_i$ is a collection of generators closed under finite intersection for the σ-algebra $\mathcal{B}_i$.

Remark 18.5. Let $(X_i)_{i\in I}$ be a family of random objects in the same probability space with X_i taking values in a measurable space $(M_i, \mathcal{B}_i)$. If M_i' is a subset of M_i containing the image of X_i and M_i' is endowed with the σ-algebra $\mathcal{B}_i' = \big\{B \cap M_i' : B \in \mathcal{B}_i\big\}$ induced by the inclusion map $M_i' \to M_i$ then each X_i can also be regarded as an $(M_i', \mathcal{B}_i')$-valued random object. Since (18.5) is equivalent to

$$\begin{aligned}\mathbb{P}\big([X_{i_1} \in B_1 \cap M_{i_1}'] \cap \ldots \cap [X_{i_n} \in B_n \cap M_{i_n}']\big)& \\ = \mathbb{P}(X_{i_1} \in B_1 \cap M_{i_1}') \cdots \mathbb{P}(X_{i_n} \in B_n \cap M_{i_n}'),&\end{aligned}$$

it follows that the family $(X_i)_{i\in I}$ is independent with each X_i being regarded as $(M_i, \mathcal{B}_i)$-valued if and only if such family is independent with each X_i being regarded as $(M_i', \mathcal{B}_i')$-valued.

If we apply measurable functions to the members of a family of independent random objects we get a new family of independent random objects. This is shown in the next result, which generalizes Proposition 15.9.

Proposition 18.6. *Consider a family $(f_i : M_i \to M_i')_{i\in I}$ of measurable maps f_i from a measurable space $(M_i, \mathcal{B}_i)$ to a measurable space $(M_i', \mathcal{B}_i')$. Let $f : \prod_{i\in I} M_i \to \prod_{i\in I} M_i'$ be defined by $f\big((x_i)_{i\in I}\big) = \big(f_i(x_i)\big)_{i\in I}$, for all $(x_i)_{i\in I} \in \prod_{i\in I} M_i$. Given, for each $i \in I$, a probability measure $\mathbb{P}_i$ on $\mathcal{B}_i$ we have:*

$$f_*\bigotimes_{i\in I}\mathbb{P}_i = \bigotimes_{i\in I}(f_i)_*\mathbb{P}_i. \tag{18.6}$$

In particular, if $(X_i)_{i\in I}$ is a family of random objects on the same probability space with X_i taking values in M_i and if $(X_i)_{i\in I}$ is independent then also $\big(f_i(X_i)\big)_{i\in I}$ is independent.

Proof. Equality (18.6) follows by noting that $f_*\bigotimes_{i\in I}\mathbb{P}_i$ satisfies the property that characterizes the product $\bigotimes_{i\in I}(f_i)_*\mathbb{P}_i$. The second part of the statement then follows from the fact that the map $\big(f_i(X_i)\big)_{i\in I}$ is equal to the composition of the map $(X_i)_{i\in I}$ with f. □

Proposition 18.6 is very easy to prove but it is not in general sufficient for handling many concrete problems. For example, one often has a situation in which, say, independent random objects X_1, X_2, X_3, X_4 are given and one wishes to conclude that $f(X_1, X_2)$ and $g(X_3, X_4)$ are independent, i.e.,

we want to aggregate our independent random objects in disjoint smaller families before applying measurable functions. It turns out that this can be obtained from Proposition 18.6, but one needs to prove first that the pair (Y_1, Y_2) of aggregated random objects $Y_1 = (X_1, X_2)$, $Y_2 = (X_3, X_4)$ is also independent. This amounts to an associative property for the product of families of probability measures that we state in full generality below.

Proposition 18.7. *Let $\big((M_i, \mathcal{B}_i, \mathbb{P}_i)\big)_{i\in I}$ be an arbitrary family of probability spaces and write $I = \bigcup_{\lambda\in\Lambda} J_\lambda$ as the union of a family $(J_\lambda)_{\lambda\in\Lambda}$ of pairwise disjoint subsets. The bijective map*

$$\prod_{\lambda\in\Lambda}\prod_{i\in J_\lambda} M_i \ni \big((x_i)_{i\in J_\lambda}\big)_{\lambda\in\Lambda} \longmapsto (x_i)_{i\in I} \in \prod_{i\in I} M_i \tag{18.7}$$

is an isomorphism of measurable spaces, where the domain is endowed with the σ-algebra $\bigotimes_{\lambda\in\Lambda}\bigotimes_{i\in J_\lambda}\mathcal{B}_i$ and the counter-domain is endowed with the σ-algebra $\bigotimes_{i\in I}\mathcal{B}_i$. Moreover, the push-forward under (18.7) *of the iterated product of probability measures*

$$\bigotimes_{\lambda\in\Lambda}\bigotimes_{i\in J_\lambda}\mathbb{P}_i \tag{18.8}$$

is equal to $\bigotimes_{i\in I}\mathbb{P}_i$.

Proof. Simply check that the push-forward of (18.8) under (18.7) satisfies the property that characterizes $\bigotimes_{i\in I}\mathbb{P}_i$. □

Corollary 18.8. *Let $(X_i)_{i\in I}$ be a family of random objects in the same probability space with X_i taking values in a measurable space $(M_i, \mathcal{B}_i)$. Write $I = \bigcup_{\lambda\in\Lambda} J_\lambda$ as the union of a family $(J_\lambda)_{\lambda\in\Lambda}$ of pairwise disjoint subsets. For each $\lambda \in \Lambda$, consider the random object $Y_\lambda = (X_i)_{i\in J_\lambda}$ taking values in $\big(\prod_{i\in J_\lambda} M_i, \bigotimes_{i\in J_\lambda}\mathcal{B}_i\big)$ whose i-th coordinate is X_i, for all $i \in J_\lambda$. The following statements are equivalent:*

- *the family $(X_i)_{i\in I}$ is independent;*
- *for every $\lambda \in \Lambda$ the family $(X_i)_{i\in J_\lambda}$ is independent and the family $(Y_\lambda)_{\lambda\in\Lambda}$ is independent.*

Proof. Follows from the fact that the map $(X_i)_{i\in I}$ is the composition of the map $(Y_\lambda)_{\lambda\in\Lambda}$ with the isomorphism (18.7). □

We now obtain the desired generalization of Proposition 18.6.

Corollary 18.9. *Let $(X_i)_{i\in I}$ be a family of random objects in the same probability space with X_i taking values in a measurable space $(M_i, \mathcal{B}_i)$. Write $I = \bigcup_{\lambda\in\Lambda} J_\lambda$ as the union of a family $(J_\lambda)_{\lambda\in\Lambda}$ of pairwise disjoint subsets. For each $\lambda \in \Lambda$ let $f_\lambda : \prod_{i\in J_\lambda} M_i \to M'_\lambda$ be a measurable map taking values in some measurable space $(M'_\lambda, \mathcal{B}'_\lambda)$ and let Y_λ be defined as in the statement of Corollary 18.8. If the family $(X_i)_{i\in I}$ is independent, then also the family $\big(f_\lambda(Y_\lambda)\big)_{\lambda\in\Lambda}$ is independent.*

Proof. Follows directly from Corollary 18.8 and Proposition 18.6. □

Let us now prove a simple useful criterion for independence of finite families of random objects in terms of probability density functions.

Proposition 18.10. *Let $(X_i)_{i=1}^n$ be an n-tuple of random objects with X_i taking values in a measurable space $(M_i, \mathcal{B}_i)$. For each $i = 1, \dots, n$, let μ_i be a nonnegative countably additive σ-finite measure defined on $\mathcal{B}_i$ and denote by μ the product measure $\bigotimes_{i=1}^n \mu_i$. If X_i admits a probability density function f_{X_i} with respect to μ_i for all $i = 1, \dots, n$ and the family $(X_i)_{i=1}^n$ is independent then the map f_X defined by*

$$f_X(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n), \quad (x_1, \dots, x_n) \in \prod_{i=1}^n M_i$$

is a probability density function for $X = (X_i)_{i=1}^n$ with respect to μ. Conversely, if X admits a probability density function with respect to μ of the form

$$f_X(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n), \quad (x_1, \dots, x_n) \in \prod_{i=1}^n M_i$$

for certain maps $f_i : M_i \to [0, +\infty[$, $i = 1, \dots, n$, then the family $(X_i)_{i=1}^n$ is independent and there are positive constants c_i, $i = 1, \dots, n$, such that $c_i f_i$ is a probability density function of X_i with respect to μ_i, for all $i = 1, \dots, n$.

Proof. For the first part, note that a simple application of Fubini–Tonelli's Theorem yields

$$\int_B f_X \, \mathrm{d}\mu = \prod_{i=1}^n \int_{B_i} f_{X_i} \, \mathrm{d}\mu_i = \prod_{i=1}^n \mathbb{P}(X_i \in B_i) = \mathbb{P}(X \in B),$$

for all $B_i \in \mathcal{B}_i$, $i = 1, \dots, n$, where $B = \prod_{i=1}^n B_i$. By Proposition 6.6, this shows that f_X is a probability density function for X with respect to μ. For the converse, we first check that each f_i is measurable. To this aim, note that since f_X cannot be identically zero, for each $i = 1, \dots, n$ there exists $\bar{x}_i \in M_i$ with $f_i(\bar{x}_i) > 0$. To conclude that f_i is measurable use that f_X is measurable and that $f_i(x_i)$ is obtained by locking the value of x_j at $\bar{x}_j$ for $j \neq i$ in $f_X(x_1, \dots, x_n)$ and by multiplying the result by a positive constant. Now an application of Fubini–Tonelli's Theorem yields

$$1 = \int_M f_X \, \mathrm{d}\mu = \prod_{i=1}^n \int_{M_i} f_i \, \mathrm{d}\mu_i$$

where $M = \prod_{i=1}^n M_i$. This implies that for each $i = 1, \dots, n$ we can find $c_i > 0$ such that $\int_{M_i} c_i f_i \, \mathrm{d}\mu_i = 1$ and $\prod_{i=1}^n c_i = 1$. We then have

$$f_X(x_1, \dots, x_n) = \prod_{i=1}^n c_i f_i(x_i)$$

and applying again Fubini-Tonelli's Theorem we get:

$$\mathbb{P}_X(B) = \int_B f_X \,\mathrm{d}\mu = \prod_{i=1}^{n} \mathbb{P}_i(B_i),$$

for all $B_i \in \mathcal{B}_i$, $i = 1, \ldots, n$, where $B = \prod_{i=1}^n B_i$ and $\mathbb{P}_i$ is the probability measure on $\mathcal{B}_i$ defined by $\mathbb{P}_i(B_i) = \int_{B_i} c_i f_i \,\mathrm{d}\mu_i$. Thus $\mathbb{P}_X = \bigotimes_{i=1}^n \mathbb{P}_i$ and taking the push-forward under the i-th projection we obtain that $\mathbb{P}_i = \mathbb{P}_{X_i}$, for all $i = 1, \ldots, n$. This implies that $c_i f_i$ is a probability density function for X_i with respect to μ_i and that the family $(X_i)_{i=1}^n$ is independent. □

18.1. **Independence of families of events.** In Section 15 we have defined independence for pairs of events. Clearly, if $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space and $A, B \in \mathcal{A}$ are events, we have that A and B are independent if and only if the indicator random variables $\mathbf{1}_A$ and $\mathbf{1}_B$ are independent. This fact suggests the following definition.

Definition 18.11. Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we say that a family of events $(A_i)_{i\in I}$ in $\mathcal{A}$ is *independent* if the family $(\mathbf{1}_{A_i})_{i\in I}$ of indicator random variables is independent.

We have the following simple characterization for independence of families of events.

Proposition 18.12. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(A_i)_{i\in I}$ be a family of events in $\mathcal{A}$. We have that $(A_i)_{i\in I}$ is independent if and only if*

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_n}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_n}),$$

for any distinct $i_1, \ldots, i_n \in I$ and any $n \geq 1$.

Proof. By Remark 18.5, we can regard each indicator random variable $\mathbf{1}_{A_i}$ as a random object taking values in the set $\{0, 1\}$ endowed with the σ-algebra $\wp\big(\{0, 1\}\big)$. Now use the fact that $\big\{\{1\}\big\}$ is a collection of generators for $\wp\big(\{0, 1\}\big)$ closed under finite intersections. □

19. Conditional expectation

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. If $A \in \mathcal{A}$ is an event with positive probability, then the map that associates to each event its conditional probability given A

$$\mathbb{P}(\,\cdot\,|A) : \mathcal{A} \ni B \longmapsto \mathbb{P}(B|A) \in [0, 1]$$

is a probability measure on Ω. Thus, if $Y : \Omega \to \mathbb{R}$ is a random variable, we can integrate Y with respect to $\mathbb{P}(\,\cdot\,|A)$ and such integral should naturally be called the *conditional expected value* (or *conditional expectation*) of Y given the event A and be denoted by $E(Y|A)$. As the measure $\mathbb{P}(\,\cdot\,|A)$ vanishes outside of A and is equal to $\frac{1}{\mathbb{P}(A)}\,\mathbb{P}$ on measurable subsets of A, we have:

$$E(Y|A) = \frac{1}{\mathbb{P}(A)} \int_A Y \,\mathrm{d}\mathbb{P}, \tag{19.1}$$

provided that the integral of Y over A with respect to $\mathbb{P}$ exists.

Just like with conditional probability, we would like to be able to condition also on events of probability zero of the form $[X = x]$, with X a random object taking values in some measurable space $(M, \mathcal{B})$. If a regular conditional probability $\mathbb{P}(\,\cdot\,|X = x)$ given X exists, we can define $E(Y|X = x)$ as the integral of Y with respect to the probability measure $\mathbb{P}(\,\cdot\,|X = x)$ or, alternatively, as the integral of the identity map of $\mathbb{R}$ with respect to the push-forward of $\mathbb{P}(\,\cdot\,|X = x)$ under Y. Such push-forward is a regular conditional probability of Y given X. Since Y is real-valued, a regular conditional probability of Y given X always exists (Theorem 17.3) so we might as well start with that and define

$$E(Y|X = x) = \int_{\mathbb{R}} y \,\mathrm{d}K(x)(y), \tag{19.2}$$

for all $x \in M$, where $K(x)(C) = \mathbb{P}(Y \in C|X = x)$ is a regular conditional probability of Y given X. Of course, the integral in (19.2) is going to depend on the choice of K but, as usual, only the equivalence class of the map $x \mapsto E(Y|X = x)$ modulo $\mathbb{P}_X$-almost sure equality is expected to be well-defined.

Now pick $B \in \mathcal{B}$ and let us integrate the map $x \mapsto E(Y|X = x)$ defined in (19.2) over B with respect to $\mathbb{P}_X$ using the generalized Fubini–Tonelli's Theorem 16.7. We have

$$\begin{aligned}\int_B E(Y|X = x)\,\mathrm{d}\mathbb{P}_X(x) &= \int_M \left(\int_{\mathbb{R}} y\mathbf{1}_B(x)\,\mathrm{d}K(x)(y) \right) \mathrm{d}\mathbb{P}_X(x) \\ &= \int_{M\times\mathbb{R}} y\mathbf{1}_B(x)\,\mathrm{d}(\mathbb{P}_X \star K)(x,y) = \int_{M\times\mathbb{R}} y\mathbf{1}_B(x)\,\mathrm{d}\mathbb{P}_{(X,Y)}(x,y) \\ &= \int_{[X\in B]} Y \,\mathrm{d}\mathbb{P},\end{aligned}$$

provided that the integral $\int_{[X\in B]} Y\,\mathrm{d}\mathbb{P}$ exists. What we have proven is that if Y is integrable then the map $x \mapsto E(Y|X = x)$ defined in (19.2) is a Radon–Nikodym derivative of the finite countably additive signed measure

$$\mathcal{B} \ni B \longmapsto \int_{[X\in B]} Y \,\mathrm{d}\mathbb{P} = E\big(Y\mathbf{1}_{[X\in B]}\big) \in \mathbb{R} \tag{19.3}$$

with respect to the probability measure $\mathbb{P}_X$. Now it is obvious that for any integrable random variable $Y : \Omega \to \mathbb{R}$ the map (19.3) defines a finite countably additive signed measure that is absolutely continuous with respect to $\mathbb{P}_X$ and thus we can give the following definition.

Definition 19.1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $Y : \Omega \to \mathbb{R}$ be a random variable with finite expected value (i.e., Y is $\mathbb{P}$-integrable) and $X : \Omega \to M$ be a random object taking values in a measurable space $(M, \mathcal{B})$. A *conditional expectation of Y given X* is any measurable real-valued map

$$M \ni x \longmapsto E(Y|X = x) \in \mathbb{R}$$

which is a Radon–Nikodym derivative of the finite countably additive signed measure (19.3) with respect to $\mathbb{P}_X$, i.e., any measurable real-valued map such that:

$$\int_B E(Y|X=x)\,\mathrm{d}\mathbb{P}_X(x) = \int_{[X\in B]} Y\,\mathrm{d}\mathbb{P} = E\big(Y\mathbf{1}_{[X\in B]}\big),$$

for any $B \in \mathcal{B}$.

The Radon–Nikodym Theorem ensures that the conditional expectation of Y given X exists and is unique up to $\mathbb{P}_X$-almost sure equality. Definition 19.1 is the standard approach for defining conditional expectation of a random variable given a random object. Using (19.2) as a definition is also possible since, as we have shown, (19.2) is consistent with Definition 19.1. Though (19.2) is more intuitive, Definition 19.1 is simpler as it has the advantage of not requiring a regular conditional probability.

Remark 19.2. If the random variable $Y : \Omega \to \mathbb{R}$ is not $\mathbb{P}$-integrable but its integral with respect to $\mathbb{P}$ at least exists in $[-\infty, +\infty]$ (equivalently, if either the positive or the negative part of Y is $\mathbb{P}$-integrable) then the countably additive signed measure (19.3) will not be finite and it might not even be σ-finite. It turns out that the Radon–Nikodym Theorem does have a version in which only the measure in the denominator is assumed to be σ-finite and the measure in the numerator is arbitrary. For such version, the Radon–Nikodym derivative might be a function taking values in the extended real line $[-\infty, +\infty]$. Thus, one can define $E(Y|X = x)$ as in Definition 19.1 assuming only that $E(Y) = \int_\Omega Y\,\mathrm{d}\mathbb{P}$ exists in $[-\infty, +\infty]$, but in that case $E(Y|X = x)$ might also take values in $[-\infty, +\infty]$.

We note that conditional probability of a fixed event $A \in \mathcal{A}$ given a random object X (as in Definition 15.4) is a particular case of conditional expectation given X. Namely, probability is a particular case of expected value, since $\mathbb{P}(A) = E(\mathbf{1}_A)$ and, similarly, $\mathbb{P}(A|X = x)$ is a particular case of $E(\mathbf{1}_A|X = x)$.

Example 19.3. Let X and Y be square integrable random variables on the same probability space, so that the covariance $\mathrm{Cov}(X, Y)$ is well-defined. Let $K(x)(B) = \mathbb{P}(Y \in B|X = x)$ be a regular conditional probability of Y given X, so that $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \star K$ and the map

$$f(x) = \int_{\mathbb{R}} y\,\mathrm{d}K(x)(y) = E(Y|X = x), \quad x \in \mathbb{R}$$

is a conditional expectation of Y given X. The expected value $E(XY)$ of the product of X and Y can be written as

$$E(XY) = \int_{\mathbb{R}^2} xy\,\mathrm{d}\mathbb{P}_{(X,Y)}(x, y) = \int_{\mathbb{R}^2} xy\,\mathrm{d}(\mathbb{P}_X \star K)(x, y)$$

and using the generalized Fubini–Tonelli's Theorem 16.7 we obtain:

$$E(XY) = \int_{\mathbb{R}} xf(x)\,\mathrm{d}\mathbb{P}_X = E\big(Xf(X)\big).$$

Moreover

$$E(Y) = \int_{\mathbb{R}} f(x)\, \mathrm{d}\mathbb{P}_X$$

so that the covariance of X and Y is given by:

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = \int_{\mathbb{R}} \big(x - E(X)\big) f(x)\, \mathrm{d}\mathbb{P}_X.$$

The latter equality says that the covariance $\mathrm{Cov}(X,Y)$ is the L^2-inner product with respect to the probability measure $\mathbb{P}_X$ between the maps $f(x) = E(Y|X = x)$ and $g(x) = x - E(X)$. The map g is clearly L^2-orthogonal to the constant maps. If X and Y are independent then the regular conditional probability K of Y given X is $\mathbb{P}_X$-almost surely constant (Proposition 15.12), so that in particular the conditional expectation $f(x) = E(Y|X = x)$ is also $\mathbb{P}_X$-almost surely constant and hence f and g are L^2-orthogonal. This is just a new proof of the fact that independent random variables have zero covariance (Corollary 15.11). However, here we get a sense of how much stronger independence is than merely zero covariance: namely, zero covariance simply says that f and g are L^2-orthogonal with respect to $\mathbb{P}_X$. While the L^2-orthogonal complement of g contains the constant maps, it is typically a very large space (usually infinite-dimensional). So $\mathrm{Cov}(X,Y) = 0$ is much weaker than the condition that the conditional expectation $f(x) = E(Y|X = x)$ be ($\mathbb{P}_X$-almost surely) independent of $x \in \mathbb{R}$ and the latter condition is much weaker than independence between X and Y, which is equivalent to the condition that the probability distribution $K(x) = \mathbb{P}(\,\cdot\,|X = x)$ of Y given $X = x$ be ($\mathbb{P}_X$-almost surely) independent of $x \in \mathbb{R}$

19.1. **Conditioning on a σ-algebra.** Let $Y : \Omega \to \mathbb{R}$ be an integrable random variable in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $X : \Omega \to M$ be a random object taking values in a measurable space $(M, \mathcal{B})$. The conditional expectation of Y given X introduced above is a measurable map $f : M \to \mathbb{R}$ and we write $f(x) = E(Y|X = x)$, for all $x \in M$.

In practical applications, we sometimes want to talk about the conditional expected value $E(Y|X = x)$ for some given fixed value of $x \in M$, but in some cases we want to think about x as a function $X(\omega)$ of the random outcome ω of the random experiment modelled by $(\Omega, \mathcal{A}, \mathbb{P})$. This change of point of view corresponds to replacing the map $f : M \to \mathbb{R}$ with the composition $f(X) = f \circ X$ of f with X, which yields a random variable $E(Y|X) = f(X) : \Omega \to \mathbb{R}$.

It follows directly from Definition 19.1 that the random variable $E(Y|X)$ satisfies

$$\int_{[X \in B]} E(Y|X)\, \mathrm{d}\mathbb{P} = \int_{[X \in B]} Y\, \mathrm{d}\mathbb{P}, \tag{19.4}$$

for every $B \in \mathcal{B}$. Moreover, $E(Y|X)$ is measurable with respect to the σ-algebra $\mathcal{A}_X = \big\{X^{-1}[B] : B \in \mathcal{B}\big\}$ induced by X, since it is a function of X.

Equality (19.4) says that $E(Y|X)$ and Y have the same integral over every element of $\mathcal{A}_X$. The fact that $E(Y|X)$ is $\mathcal{A}_X$-measurable taken together with (19.4) means that $E(Y|X)$ is a Radon–Nikodym derivative of the finite countably additive signed measure

$$\mathcal{A}_X \ni A \longmapsto \int_A Y \, \mathrm{d}\mathbb{P} \in \mathbb{R} \tag{19.5}$$

with respect to the probability measure on $\mathcal{A}_X$ given by the restriction of $\mathbb{P}$. It is obvious that (19.5) is indeed absolutely continuous with respect to the restriction of $\mathbb{P}$ to $\mathcal{A}_X$. What this shows is that we can give a description of the conditional expectation $E(Y|X)$ without any direct reference to the random object X, but instead using only the σ-algebra $\mathcal{A}_X$ induced by X. This leads naturally to the following definition.

Definition 19.4. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $Y : \Omega \to \mathbb{R}$ be an integrable random variable and $\mathcal{B}$ be a σ-algebra of subsets of Ω contained in $\mathcal{A}$. We define a *conditional expectation of Y given $\mathcal{B}$*, denoted $E(Y|\mathcal{B})$, as a Radon–Nikodym derivative of the finite countably additive signed measure

$$\mathcal{B} \ni A \longmapsto \int_A Y \, \mathrm{d}\mathbb{P} = E(Y\mathbf{1}_A) \in \mathbb{R}$$

with respect to the restriction of $\mathbb{P}$ to $\mathcal{A}$, i.e., $E(Y|\mathcal{B}) : \Omega \to \mathbb{R}$ is a $\mathcal{B}$-measurable map which has the same integral as Y with respect to $\mathbb{P}$ over any $A \in \mathcal{B}$.

The Radon–Nikodym Theorem ensures that the conditional expectation of Y given $\mathcal{B}$ exists and is unique up to $\mathbb{P}$-almost sure equality. As discussed above, the conditional expectation of Y given the σ-algebra $\mathcal{A}_X$ induced by a random object X is the same thing as $E(Y|X)$, i.e., the random variable obtained by composing $x \mapsto E(Y|X = x)$ with X.

We note that, as in the case of conditioning on a random object, the assumption that Y be integrable in Definition 19.4 can be relaxed to the assumption that $E(Y)$ exists in $[-\infty, +\infty]$, as long as one allows $E(Y|\mathcal{B})$ to take values in $[-\infty, +\infty]$ (see Remark 19.2).

It turns out that the map $x \mapsto E(Y|X = x)$ (or, more precisely, its class of maps modulo $\mathbb{P}_X$-almost everywhere equality) can be recovered from $E(Y|X) = E(Y|\mathcal{A}_X)$.

Lemma 19.5. *Let Ω be a set, $(M, \mathcal{B})$ be a measurable space, $X : \Omega \to M$ be a map and denote by $\mathcal{A}_X$ the σ-algebra of subsets of Ω induced by X. A real-valued map $g : \Omega \to \mathbb{R}$ is measurable with respect to $\mathcal{A}_X$ if and only if there exists a measurable map $f : M \to \mathbb{R}$ such that $g = f \circ X$.*

Proof. The nontrivial part of the thesis is that if g is $\mathcal{A}_X$-measurable then $g = f \circ X$ for some measurable map $f : M \to \mathbb{R}$. The validity of such statement is clear if g is an indicator function or a simple map, i.e., a finite linear combination of indicator functions. For the general case, note that if $g : \Omega \to \mathbb{R}$ is $\mathcal{A}_X$-measurable then g is the pointwise limit of a sequence

$(g_n)_{n\geq 1}$ of simple $\mathcal{A}_X$-measurable maps $g_n : \Omega \to \mathbb{R}$ and thus for each $n \geq 1$ we have $g_n = f_n \circ X$ for some measurable map $f_n : M \to \mathbb{R}$. The conclusion is obtained by taking f for instance as the lim sup of $(f_n)_{n\geq 1}$ (replacing $\pm\infty$ values that f might assume outside the image of X by some arbitrary finite value). □

Using Lemma 19.5 we see that if Y in an integrable random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, $X : \Omega \to M$ is a random object taking values in a measurable space $(M, \mathcal{B})$ and $E(Y|\mathcal{A}_X)$ is a conditional expectation of Y given the σ-algebra $\mathcal{A}_X$ induced by X then $E(Y|\mathcal{A}_X) = f(X)$ for some measurable map $f : M \to \mathbb{R}$. The map f then satisfies

$$\int_B f \, \mathrm{d}\mathbb{P}_X = \int_{[X\in B]} f(X) \, \mathrm{d}\mathbb{P} = \int_{[X\in B]} Y \, \mathrm{d}\mathbb{P}$$

for all $B \in \mathcal{B}$ and thus it is a valid conditional expectation of Y given X in the sense of Definition 19.1.

We observe that a conditional expectation $E(Y|\mathcal{B})$ on an arbitrary σ-algebra $\mathcal{B}$ contained in $\mathcal{A}$ can always be seen as a particular case of $E(Y|X)$ for some random object X. For example, take X as the identity map from $(\Omega, \mathcal{A})$ to $(\Omega, \mathcal{B})$ so that $\mathcal{A}_X = \mathcal{B}$.

The idea of conditioning on σ-algebras might at first sound weird as one normally thinks that "conditioning" should always mean "conditioning on some fact", such as the fact that an event like $[X = x]$ happened. To make conditioning on a σ-algebra more palatable, one should think about a σ-algebra $\mathcal{B} \subset \mathcal{A}$ as representing a certain amount of information regarding the outcome $\omega \in \Omega$. More specifically, in nonpathological cases (such as if $\mathcal{B}$ is induced by a random object taking values on a separable metric space endowed with its Borel σ-algebra — see the discussion in Remark 15.6), one can think of an agent as being $\mathcal{B}$-*informed* if such agent knows enough about $\omega \in \Omega$ to figure out whether or not ω is in B, for any $B \in \mathcal{B}$. Alternatively, a $\mathcal{B}$-informed agent is an agent that knows the value of any $\mathcal{B}$-measurable random variable.

Larger σ-algebras correspond to more information. For example, the largest possible σ-algebra, which is the σ-algebra $\mathcal{A}$ in which the probability measure is defined, should be thought as the perfect information σ-algebra, i.e., $\mathcal{A}$-informed agents know everything. On the other extreme, the smallest possible σ-algebra, which is just $\{\emptyset, \Omega\}$, corresponds to no information at all. If $Y : \Omega \to \mathbb{R}$ is an integrable random variable, then the expected value of Y conditioned on the no-information σ-algebra $\{\emptyset, \Omega\}$ is simply the standard expected value $E(Y)$ of Y — or, more precisely, the random variable that is constant and equal to $E(Y)$. That makes sense, as conditioning on no-information is the same as not conditioning on anything. On the other extreme, the conditional expectation $E(Y|\mathcal{A})$ of Y on the perfect information σ-algebra $\mathcal{A}$ is just Y itself. This also makes sense, as the $\mathcal{A}$-informed agent knows the true value of $\omega \in \Omega$ and for such agent the expected value of Y is the true value $Y(\omega)$ of Y.

To get further insight into this matter let us look at the case of a σ-algebra generated by a countable partition. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and suppose we are given a countable partition of Ω into measurable subsets. We can represent such partition by the corresponding equivalence relation $\sim$ on Ω, which is an equivalence relation whose equivalence classes are in $\mathcal{A}$ and such that the quotient set $\Omega/\sim$ is countable. Let $\mathcal{B}$ be the σ-algebra of subsets of Ω generated by the collection of all equivalence classes. Clearly, $\mathcal{B}$ coincides with the collection of all possible unions of equivalence classes, which is also the σ-algebra induced by the quotient map $q : \Omega \to \Omega/\sim$, where $\Omega/\sim$ is endowed with the σ-algebra $\wp(\Omega/\sim)$ of all its subsets. The equivalence classes are then what are usually called *atoms* of $\mathcal{B}$, i.e., an equivalence class A is a nonempty element $A \in \mathcal{B}$ such that the only subsets of A that are in $\mathcal{B}$ are the empty set and A itself.

A $\mathcal{B}$-measurable random variable is simply a random variable that is constant on every atom. Given an integrable random variable $Y : \Omega \to \mathbb{R}$, it is easily checked that $E(Y|\mathcal{B})$ is a random variable whose restriction to every positive probability atom A of $\mathcal{B}$ is constant and equal to the conditional expectation $E(Y|A)$ of Y on the event A (recall (19.1)). Thus, taking the conditional expectation of Y on the σ-algebra $\mathcal{B}$ has the effect of forcing Y to become constant on the atoms of $\mathcal{B}$ by averaging Y inside such atoms. A $\mathcal{B}$-informed agent does not know the true value of $\omega \in \Omega$, but knows what is the atom A of $\mathcal{B}$ that contains ω. Thus, for such an agent, the expected value of Y is the expected value of Y conditioned on A.

20. Images of densities under diffeomorphisms

If X is an $\mathbb{R}^n$-valued random vector which admits a probability density function $f_X : \mathbb{R}^n \to [0, +\infty[$ (with respect to Lebesgue measure) it would be useful to be able to obtain a formula for the probability density function of a random vector of the form $\phi(X)$, where $\phi : \mathbb{R}^n \to \mathbb{R}^m$ is a sufficiently nice function (say, a function of class C^1). However, it is not true in general that $\phi(X)$ admits a probability density function, even when ϕ is really nice. For instance, if ϕ is constant, then the distribution of $\phi(X)$ is a Dirac delta (Definition 13.1) which is not absolutely continuous with respect to Lebesgue measure. More generally, if the image of ϕ has null Lebesgue measure, then $\mathbb{P}_{\phi(X)}$ cannot be absolutely continuous with respect to Lebesgue measure as the $\mathbb{P}_{\phi(X)}$-probability of a Borel set that contains the image of ϕ is equal to one. In this section we will study the case in which ϕ is a local diffeomorphism of class C^1.

Recall that a *diffeomorphism* $\phi : U \to V$ between open subsets U and V of $\mathbb{R}^n$ is a bijective differentiable map whose inverse is also differentiable. If ϕ is of class C^k (with $1 \leq k \leq +\infty$) and ϕ^{-1} is differentiable then ϕ^{-1} is automatically of class C^k. A map $\phi : U \to \mathbb{R}^n$ defined in an open subset U of $\mathbb{R}^n$ is said to be a *local diffeomorphism* if every point of U has an open neighborhood in U such that the restriction of ϕ to such neighborhood is a

diffeomorphism onto some open subset of $\mathbb{R}^n$. If ϕ is a local diffeomorphism then the image of ϕ is open and if ϕ is also injective then ϕ is actually a diffeomorphism onto its image.

For a differentiable map $\phi : U \to \mathbb{R}^m$ defined in an open subset U of $\mathbb{R}^n$, we denote by $\mathrm{d}\phi(x)$ its differential at the point $x \in U$, which is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. The matrix that represents $\mathrm{d}\phi(x)$ with respect to the canonical bases is the so called *Jacobian matrix* of ϕ at x. The celebrated *Inverse Function Theorem* states that if $\phi : U \to \mathbb{R}^n$ is of class C^1 then ϕ is a local diffeomorphism if and only if $\mathrm{d}\phi(x)$ is an isomorphism (equivalently, the determinant of $\mathrm{d}\phi(x)$ is nonzero) for every $x \in U$.

The key ingredient for obtaining a probability density function for $\phi(X)$ is the change of variables theorem for Lebesgue integration in $\mathbb{R}^n$ which we recall below. We will denote by $\mathfrak{m}$ the Lebesgue measure of $\mathbb{R}^n$ and all probability density functions of $\mathbb{R}^n$-valued random vectors will be taken with respect to (the restriction to the Borel σ-algebra of) $\mathfrak{m}$.

Integration theorems will be stated with more generality than what is needed for the probability theory related applications, for the sake of completeness. We will call a function f defined in some subset of $\mathbb{R}^n$ *Lebesgue-measurable* if it is measurable with respect to the Lebesgue σ-algebra on its domain and *Borel-measurable* if it is measurable with respect to the Borel σ-algebra on its domain, where the Borel σ-algebra is used for the counter-domain in both cases. Recall that a map of class C^1 from an open subset of $\mathbb{R}^n$ to $\mathbb{R}^n$, being locally Lipschitz, maps sets of Lebesgue measure zero to sets of Lebesgue measure zero. Thus C^1 diffeomorphisms map Lebesgue measurable sets to Lebesgue measurable sets.

Theorem 20.1. *Let $\phi : U \to V$ be a diffeomorphism of class C^1 between open subsets U and V of $\mathbb{R}^n$. If $f : V \to [-\infty, +\infty]$ is a Lebesgue-measurable function then*

$$\int_U f\big(\phi(x)\big)\big|\det\big(\mathrm{d}\phi(x)\big)\big|\,\mathrm{d}\mathfrak{m}(x) = \int_V f(y)\,\mathrm{d}\mathfrak{m}(y), \tag{20.1}$$

meaning that the integral on the lefthand side of the equality exists if and only if the integral on the righthand side of the equality exists and that they are equal when both exist. □

Theorem 20.1 could be restated by saying that the push-forward under ϕ of the measure on U given by integration of $x \mapsto \big|\det\big(\mathrm{d}\phi(x)\big)\big|$ with respect to Lebesgue measure is the Lebesgue measure on V. Equality (20.1) then follows from the simple abstract version of the change of variables theorem given in Proposition 9.1.

Remark 20.2. Though we will not need such generalization, we mention that Theorem 20.1 also holds if $\phi : U \to \mathbb{R}^n$ is merely an injective function of class C^1 defined in an open subset U of $\mathbb{R}^n$ and V is the image of ϕ. In this case V might not be open, but it is Borel since it is a countable union of

compact sets. To obtain such generalization, note that if U_0 is the (open) set of points $x \in U$ with $\det\big(\mathrm{d}\phi(x)\big) \neq 0$ and $V_0 = \phi[U_0]$ then Theorem 20.1 can be applied to the restriction of ϕ to U_0, which is a diffeomorphism onto V_0. Moreover, the integrand on the lefthand side of (20.1) obviously vanishes on $U \setminus U_0$ and the set $V \setminus V_0$ is contained in the set of critical values of ϕ and thus has Lebesgue measure zero due to Sard's Theorem.

As a consequence of Theorem 20.1 we immediately obtain the following method for obtaining a probability density function for $\phi(X)$ if ϕ is a diffeomorphism of class C^1 and X is a random vector admitting a probability density function.

Proposition 20.3. *Let $\phi : U \to V$ be a diffeomorphism of class C^1 between open subsets U and V of $\mathbb{R}^n$ and let X be an $\mathbb{R}^n$-valued random vector with image contained in U (so that X can be regarded as an U-valued random object). If X admits a probability density function $f_X : U \to [0, +\infty[$ then a probability density function $f_{\phi(X)} : V \to [0, +\infty[$ for $\phi(X)$ is given by:*

$$f_{\phi(X)}(y) = \frac{f_X(x)}{\big|\det\big(\mathrm{d}\phi(x)\big)\big|}, \tag{20.2}$$

for all $y \in V$, where $x = \phi^{-1}(y)$.

Proof. Apply Theorem 20.1 with f being the function defined by (20.2) multiplied by the indicator function of a Borel subset of V. □

In the statement of Proposition 20.3 the function $f_{\phi(X)}$ is defined only on V, so it is a probability density function for $\phi(X)$ regarded as a V-valued random object. If we want a probability density function for $\phi(X)$ regarded as an $\mathbb{R}^n$-valued random vector, we can just extend $f_{\phi(X)}$ to all of $\mathbb{R}^n$ by assiging to it the value zero outside of V.

Example 20.4. Let X and Y be $\mathbb{R}^n$-valued random vectors on the same probability space and assume that they admit a joint probability density function $f_{(X,Y)} : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty[$. Let us obtain a probability density function for $X+Y$ using Proposition 20.3. We cannot apply the proposition directly with ϕ being the sum map from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^n$, but we can use it with the diffeomorphism $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ given by

$$\phi(x, y) = (x + y, y),$$

for all $x, y \in \mathbb{R}^n$. Note that ϕ is linear, so that its differential $\mathrm{d}\phi(x, y)$ at any point $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ is equal to the linear map ϕ itself. Moreover, $\det \phi = 1$, as the matrix of ϕ has a null lower left $n \times n$ block and two $n \times n$ identity blocks over the main diagonal. Proposition 20.3 then yields

$$f_{(X+Y,Y)}(u, y) = f_{(X,Y)}\big(\phi^{-1}(u, y)\big) = f_{(X,Y)}(u - y, y),$$

for all $(u, y) \in \mathbb{R}^n \times \mathbb{R}^n$, where $f_{(X+Y,Y)}$ denotes a probability density function for $(X + Y, Y) = \phi(X, Y)$. A probability density function f_{X+Y} for

$X + Y$ can now be obtained by simply integrating over y (see Example 8.4)

$$f_{X+Y}(u) = \int_{\mathbb{R}^n} f_{(X,Y)}(u - y, y)\, \mathrm{d}\mathfrak{m}(y), \quad u \in \mathbb{R}^n,$$

taking care of replacing the infinite values that f_{X+Y} might assume in a set of null measure with some fixed finite value. In the particular case in which X and Y are independent with probability density functions f_X and f_Y, we have $f_{(X,Y)}(x, y) = f_X(x) f_Y(y)$ (Proposition 18.10) and thus

$$f_{X+Y}(u) = \int_{\mathbb{R}^n} f_X(u - y) f_Y(y)\, \mathrm{d}\mathfrak{m}(y), \quad u \in \mathbb{R}^n,$$

i.e., f_{X+Y} is ($\mathfrak{m}$-almost everywhere equal to) the *convolution* $f_X * f_Y$ of f_X and f_Y.

The assumption in Proposition 20.3 that the map ϕ be injective is too strong and we need to get rid of it. For this we need a better version of Theorem 20.1.

Theorem 20.5. *Let $\phi : U \to \mathbb{R}^n$ be a local diffeomorphism of class C^1 defined in an open subset U of $\mathbb{R}^n$ and let V be the (automatically open) image of ϕ. Assume that $g : U \to [-\infty, +\infty]$ is a Lebesgue-measurable function whose integral over U with respect to the Lebesgue measure exists. The map $f : V \to [-\infty, +\infty]$ given by*

$$f(y) = \sum_{x \in \phi^{-1}(y)} \frac{g(x)}{\left|\det\big(\mathrm{d}\phi(x)\big)\right|}, \quad y \in V \tag{20.3}$$

is well-defined for Lebesgue-almost every $y \in V$. The map f is always Lebesgue-measurable and it is also Borel-measurable if g is Borel-measurable. Moreover, the integral of f over V with respect to the Lebesgue measure exists and it is equal to the integral of g over U with respect to the Lebesgue measure.

In the statement of the theorem above, when we say that "$f(y)$ is well-defined" for a certain $y \in V$ what we mean is that either the sum of the positive parts or the sum of the negative parts of the terms in the sum appearing in (20.3) is finite. In order to make f a Lebesgue-measurable function on all of V one has simply to choose some arbitrary value for $f(y)$ when it is not well-defined. When g is Borel-measurable and we want f to be Borel-measurable on all of V, we have to make the choice of values for $f(y)$ in the undefined cases in a Borel-measurable way (say, by picking a constant value).

Proof of Theorem 20.5. We can assume that g is nonnegative, as the general case will then follow by applying the result to the positive and negative part of g. If there exists an open subset U_0 of U such that ϕ is injective on U_0 and g vanishes on $U \setminus U_0$ then the result follows directly from Theorem 20.1. In general, we can cover U by countably many open sets in which ϕ is injective

and by the standard disjointification method we write U as a disjoint union $\bigcup_{n=1}^{\infty} B_n$ of Borel subsets B_n such that each B_n is contained in some open subset of U in which ϕ is injective. The conclusion then follows by noting that $g = \sum_{n=1}^{\infty} g\mathbf{1}_{B_n}$ and that the result holds for each $g\mathbf{1}_{B_n}$. □

As a corollary of Theorem 20.5 we obtain the improved version of Proposition 20.3 in which ϕ is not assumed to be injective. We will also allow for the assumptions of Proposition 20.3 to fail on sets of measure zero. In order to simplify the statement, we adopt the convention that the sum of an empty family is equal to zero.

Proposition 20.6. *Let $\phi : U \to \mathbb{R}^n$ be a map of class C^1 defined in some open subset U of $\mathbb{R}^n$ and assume that $\det\big(\mathrm{d}\phi(x)\big) \neq 0$ for Lebesgue-almost every $x \in U$. Let X be an $\mathbb{R}^n$-valued random vector with $\mathbb{P}(X \in U) = 1$ and consider the $\mathbb{R}^n$-valued random vector $\phi(X)$, to which we assign some arbitrary fixed value in the probability zero set $[X \notin U]$. If X admits a probability density function $f_X : \mathbb{R}^n \to [0, +\infty[$ then a probability density function $f_{\phi(X)} : \mathbb{R}^n \to [0, +\infty[$ for $\phi(X)$ exists and it is given by*

$$f_{\phi(X)}(y) = \sum_{x \in \phi^{-1}(y)} \frac{f_X(x)}{\big|\det\big(\mathrm{d}\phi(x)\big)\big|},$$

for all $y \in \mathbb{R}^n$, except for y in some subset of $\mathbb{R}^n$ with null Lebesgue measure in which the sum above is infinite or any of the denominators appearing in the sum vanish (in which case we replace the sum with some arbitrary fixed nonnegative finite value).

Proof. If $\det\big(\mathrm{d}\phi(x)\big) \neq 0$ for all $x \in U$ and the image of X is contained in U, the result follows by picking an arbitrary Borel subset B of $\mathbb{R}^n$ and applying Theorem 20.5 with g equal to the restriction of $f_X \mathbf{1}_{\phi^{-1}[B]}$ to U, keeping in mind that $f_{\phi(X)}$ vanishes outside of $V = \phi[U]$ and f_X vanishes Lebesgue-almost everywhere outside of U. For the general case, set

$$U_0 = \big\{x \in U : \det\big(\mathrm{d}\phi(x)\big) \neq 0\big\},$$

so that U_0 is open and $U \setminus U_0$ has Lebesgue measure zero. Since $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure, we have that $\mathbb{P}(X \in U_0) = 1$ and thus we can change X in a set of measure zero so that its image becomes contained in U_0. Such adjustment does not alter the distributions of X and $\phi(X)$. Applying the version of the proposition which we have already proven with $\phi|_{U_0}$ in the place of ϕ, we obtain that

$$f_{\phi(X)}(y) = \sum_{x \in \phi^{-1}(y) \cap U_0} \frac{f_X(x)}{\big|\det\big(\mathrm{d}\phi(x)\big)\big|}, \quad y \in \mathbb{R}^n$$

defines a probability density function for $\phi(X)$. The conclusion then follows by noting that for $y \in \mathbb{R}^n$ outside of the null Lebesgue measure set $\phi[U \setminus U_0]$ we have that $\phi^{-1}(y) = \phi^{-1}(y) \cap U_0$. □

21. The univariate normal distribution

The normal distribution is one of the most basic and important probability distributions in statistics. A standard motivation for studying such distribution is the celebrated Central Limit Theorem. In this section we will simply present the definition of the univariate normal distribution. The multivariate case will be discussed later in Section 24.

When we talk about "defining a distribution" (such as the normal distribution) what we mean is that we are going to define a certain collection of probability measures on the real line endowed with the Borel σ-algebra (or a collection of probability measures on some other measurable space). It is often convenient to define such collection using the language of random variables (or random objects), i.e., we define that a random variable X has a certain type of distribution under certain conditions. Yet, what really matters is the collection of probability measures $\mathbb{P}_X$ being defined by that statement, i.e., saying that X has a certain distribution means that $\mathbb{P}_X$ belongs to a certain collection of probability measures. The map X itself is not important unless one is interested in discussing the relationship between X and other random objects defined on the same probability space. Note that given a probability measure $\mathbb{P}$ on the real line (resp., on some other measurable space) one can always obtain a random variable (resp., random object) X with $\mathbb{P}_X = \mathbb{P}$ by letting X be the identity map, where the domain of X is endowed with the probability measure $\mathbb{P}$.

We will denote by $\mathfrak{m}$ the Lebesgue measure of $\mathbb{R}$ (restricted to the Borel σ-algebra) and, as usual, probability density functions of random variables will be taken with respect to $\mathfrak{m}$. We will say that a random variable has a (nondegenerate) normal distribution if it admits a probability density function that is the exponential of a second degree polynomial. Clearly, the exponential of a second degree polynomial has finite integral if and only if the leading coefficient is negative and thus, since the integral of a probability density function must be equal to 1, only second degree polynomials with a negative leading coefficient are admissible.

Definition 21.1. A random variable X is said to have a *nondegenerate normal distribution* (alternatively, X is a *nondegenerate normal random variable* or $\mathbb{P}_X$ is a *nondegenerate normal distribution* on $\mathbb{R}$) if

$$f_X(x) = e^{-ax^2+bx+c}, \quad x \in \mathbb{R} \tag{21.1}$$

is a probability density function for X for certain $a, b, c \in \mathbb{R}$ with $a > 0$.

We used above the name "nondegenerate normal distribution" instead of just "normal distribution" because random variables whose distribution is a Dirac delta (i.e., almost surely constant random variables) will also be regarded as normal — that is the "degenerate" case. Most textbooks simply use something equivalent to Definition 21.1 above as their definition of normal distribution, but we find convenient to include also the degenerate case under the umbrella of "normal".

Definition 21.2. A random variable is said to have a *normal distribution* (alternatively, X is a *normal random variable* or $\mathbb{P}_X$ is a *normal distribution* on $\mathbb{R}$) if either X has a nondegenerate normal distribution or X is almost surely constant.

Note that the fact that f_X must have integral equal to 1 implies that the constant term c of the polynomial in (21.1) is determined by a and b, as the following equality must hold:

$$e^{-c} = \int_{-\infty}^{+\infty} e^{-ax^2+bx} \, \mathrm{d}\mathfrak{m}(x).$$

Proposition 21.3. *If X is a normal random variable then $\alpha X + \beta$ is also a normal random variable, for any $\alpha, \beta \in \mathbb{R}$.*

Proof. If either X is almost surely constant or $\alpha = 0$ then $\alpha X + \beta$ is almost surely constant. Moreover, if X admits a probability density function given by (21.1) and $\alpha \neq 0$ then, using the formula given in Example 8.8, we see that $\alpha X + \beta$ admits a probability density function given by

$$f_{\alpha X+\beta}(y) = \frac{1}{|\alpha|} f_X\Big(\frac{y-\beta}{\alpha}\Big), \quad y \in \mathbb{R}.$$

Such probability density function is given by the exponential of the second degree polynomial

$$-\frac{a}{\alpha^2}(y-\beta)^2 + \frac{b}{\alpha}(y-\beta) + c - \ln|\alpha|$$

which has a negative leading coefficient. □

Let us derive a formula for the probability density function of a nondegenerate normal random variable that is more useful for practical applications as it explicitly shows the expected value and variance. We start by considering the integral of the exponential of the simplest second degree polynomial with a negative leading coefficient:

$$\int_{-\infty}^{+\infty} e^{-x^2} \, \mathrm{d}\mathfrak{m}(x).$$

There is a standard trick from basic multivariate calculus courses for computing this integral, which is to write its square as a double integral and then switch to polar coordinates. One then obtains that the integral is equal to $\sqrt{\pi}$ and therefore

$$f_X(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}, \quad x \in \mathbb{R}$$

is the probability density function of a random variable X. Note that f_X is of the form (21.1) as the multiplicative constant can be turned into an additive constant in the exponent. Let us compute the expected value and variance of X. The expected value is given by (recall Example 9.2)

$$E(X) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x e^{-x^2} \, \mathrm{d}\mathfrak{m}(x)$$

and since (the integral is finite and) the integrand is an odd function we get $E(X) = 0$. The variance of X is given by:

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = E(X^2) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x^2 e^{-x^2} \, \mathrm{d}\mathfrak{m}(x).$$

One then easily obtains that $\mathrm{Var}(X) = \frac{1}{2}$ by writing $x^2 e^{-x^2} = x(xe^{-x^2})$ and by computing the latter integral using integration by parts. Thus, if we set $Z = \sqrt{2}\, X$, we have that Z is a normal random variable with $E(Z) = 0$ and $\mathrm{Var}(Z) = 1$. Moreover, a probability density function f_Z for Z is given by (recall Example 8.8):

$$f_Z(z) = \frac{1}{\sqrt{2}} f_X\Big(\frac{z}{\sqrt{2}}\Big) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad z \in \mathbb{R}.$$

Definition 21.4. A random variable Z is said to have a *standard normal distribution* (alternatively, Z is a *standard normal random variable* or $\mathbb{P}_Z$ is a *standard normal distribution* on $\mathbb{R}$) if

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad z \in \mathbb{R} \tag{21.2}$$

is a probability density function for Z.

Since $E(Z) = 0$ and $\mathrm{Var}(Z) = 1$, given $\mu \in \mathbb{R}$ and $\sigma \in \,]0, +\infty[$, we have that $\mu + \sigma Z$ is a normal random variable with expected value μ and variance σ^2. A probability density function for $\mu + \sigma Z$ is given by:

$$f_{\mu+\sigma Z}(x) = \frac{1}{\sigma} f_Z\Big(\frac{x-\mu}{\sigma}\Big) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}. \tag{21.3}$$

We have shown that (21.3) is a probability density function for a nondegenerate normal random variable with mean μ and variance σ^2. Let us verify that, conversely, every nondegenerate normal random variable with mean μ and variance σ^2 has (21.3) as a probability density function. Let then X be a nondegenerate normal random variable with probability density function given by (21.1). We can choose $\mu \in \mathbb{R}$ and $\sigma > 0$ such that the polynomial in the exponent in (21.1) is equal to the polynomial in the exponent in (21.3) up to the independent term. Namely, $\mu \in \mathbb{R}$ and $\sigma > 0$ must be chosen such that

$$a = \frac{1}{2\sigma^2}, \quad b = \frac{\mu}{\sigma^2} \tag{21.4}$$

and such equalities are equivalent to:

$$\mu = \frac{b}{2a}, \quad \sigma = \frac{1}{\sqrt{2a}}. \tag{21.5}$$

Thus, defining μ and σ by (21.5) we obtain

$$e^{-ax^2+bx+c} = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{21.6}$$

for all $x \in \mathbb{R}$, since the independent term of the polynomial in the exponent of a probability density function of the form (21.1) is determined by the two other coefficients of the polynomial. Equality (21.6) implies that, if μ and σ are defined by (21.5), then μ and σ^2 are indeed the expected value and the variance, respectively, of the normal random variable X with probability density function (21.1).

We have proven the following result.

Proposition 21.5. *Every normal random variable has a finite expected value and a finite variance. Moreover, given $\mu \in \mathbb{R}$ and $\sigma > 0$, we have that a normal random variable X has expected value μ and variance σ^2 if and only if*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

is a probability density function for X. In particular, two normal random variables with the same expected value and the same variance have the same distribution. □

The fact that a normal distribution is completely determined by the expected value and the variance makes the following definition useful.

Definition 21.6. Given $\mu \in \mathbb{R}$ and $\sigma \geq 0$, we write

$$X \sim N(\mu, \sigma^2)$$

to indicate that X is a normal random variable with expected value μ and variance σ^2. In this case we also say that $\mathbb{P}_X$ is a *normal distribution with mean μ and variance σ^2* (or a *normal distribution with mean μ and standard deviation σ*).

Clearly $X \sim N(\mu, 0)$ if and only if $X = \mu$ almost surely, i.e., if and only if $\mathbb{P}_X$ is a Dirac delta centered at μ.

Observe that the univariate normal distribution can be thought as a collection of probability measures on the Borel σ-algebra of the real line and also as a family of probability measures on Borel σ-algebra of the real line indexed (in a one-to-one manner) by the parameters $\mu \in \mathbb{R}$ and $\sigma \geq 0$. Though μ is the expected value of a random variable X with $X \sim N(\mu, \sigma^2)$, it is more usual to call μ the *mean* in this context since here we are thinking about μ as a property of a probability distribution.

22. The uniform distribution

The uniform distribution models the experiment of choosing an element from a set in a such a way that every element of that set has the same probability of being chosen. This is the exact definition of the uniform distribution in the discrete case, which we consider first. For the general case this can be regarded as an informal description which motivates the precise definition.

Definition 22.1. Let M be a nonempty finite set endowed with the σ-algebra $\wp(M)$ of all its subsets. The *discrete uniform distribuiton* on M is the unique probability measure $\mathbb{P} : \wp(M) \to [0,1]$ which assigns the same probability to all points of M. More explicitly, $\mathbb{P}$ is given by

$$\mathbb{P}(A) = \frac{|A|}{|M|},$$

where $|\cdot|$ denotes the number of elements of a set. An $\big(M, \wp(M)\big)$-valued random object X is said to have a *discrete uniform distribution* if $\mathbb{P}_X$ is a discrete uniform distribution.

Note that if M is infinite and $\mathcal{B}$ is a σ-algebra of subsets of M that contains all singletons then a probability measure on $\mathcal{B}$ that assigns the same probability to every point of M must assign a null probability to every point. It follows that if such a probability measure exists then M must be uncountable, otherwise the probability of M itself would be zero.

Unlike the finite case, if M is uncountable then merely requiring that all points of M have the same probability does not adequately express the idea of uniformity. For instance, if M is a measurable subset of the real line, then every probability measure that is absolutely continuous with respect to the Lebesgue measure assigns a null probability to every point. In this context, uniformity is properly expressed by requiring the probability density function with respect to the Lebesgue measure to be constant. This leads us to the definition below.

Definition 22.2. Let M be a Borel subset of $\mathbb{R}^n$ endowed with its Borel σ-algebra $\mathcal{B}$. If $0 < \mathfrak{m}(M) < +\infty$ then the *uniform distribution* on M is the probability measure $\mathbb{P} : \mathcal{B} \to [0,1]$ defined by

$$\mathbb{P}(B) = \frac{\mathfrak{m}(B)}{\mathfrak{m}(M)},$$

for all $B \in \mathcal{B}$, where $\mathfrak{m}$ denotes the Lebesgue measure of $\mathbb{R}^n$. An $(M, \mathcal{B})$-valued random object X is said to have a *uniform distribution* if $\mathbb{P}_X$ is a uniform distribution.

Example 22.3. If a random variable X has a uniform distribution on the interval $[a,b]$, with $a < b$, then its expected value

$$E(X) = \frac{1}{b-a}\int_a^b x \, \mathrm{d}\mathfrak{m}(x) = \frac{a+b}{2}$$

is the midpoint of the interval $[a,b]$. The expected value of X^2 is given by

$$E(X^2) = \frac{1}{b-a}\int_a^b x^2 \, \mathrm{d}\mathfrak{m}(x) = \frac{1}{3}(a^2 + ab + b^2)$$

and hence the variance $\mathrm{Var}(X) = E(X^2) - E(X)^2$ is equal to:

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

More generally, if X has a uniform distribution on a Borel subset M of $\mathbb{R}^n$ with $0 < \mathfrak{m}(M) < +\infty$ then the expected value of X (if it exists) is what is normally called the *center of mass* of M.

22.1. **Abstract generalization.** A natural way to generalize the concept of uniform distribution to more abstract settings is through the theory of invariant measures. We recall some relevant definitions. Let X be a locally compact Hausdorff topological space. A *nonnegative regular Borel measure* μ on X is a nonnegative countably additive measure μ defined on the Borel σ-algebra of X satisfying the following conditions:

(i) $\mu(B)$ is the infimum of $\big\{\mu(U) : U \supset B \text{ open}\big\}$, for every Borel subset B of X;
(ii) $\mu(U)$ is the supremum of $\big\{\mu(K) : K \subset U \text{ compact}\big\}$, for every open subset U of X;
(iii) $\mu(K)$ is finite for every compact subset K of X.

It is well-known (see [4, Theorem 2.14]) that nonnegative regular Borel measures on X are in one-to-one correspondence with positive linear functionals on the space of continuous real-valued functions on X with compact support. Such correspondence associates to every nonnegative regular Borel measure μ the integration functional $f \mapsto \int_X f \, \mathrm{d}\mu$.

Let G be a locally compact Hausdorff topological group and assume that a locally compact Hausdorff topological space X is endowed with a continuous transitive action $G \times X \ni (g, x) \mapsto g \cdot x \in X$ such that the mapping $G \ni g \mapsto g \cdot x \in X$ is open for some (and hence for all) $x \in X$. A nonnegative regular Borel measure μ on X is said to be *G-invariant* if $\mu(g \cdot B) = \mu(B)$, for every $g \in G$ and every Borel subset B of X, where $g \cdot B = \big\{g \cdot x : x \in B\big\}$. It is well-known ([2, Theorem 2.7.11]) that if μ is a nonzero nonnegative G-invariant regular Borel measure on X then every other nonnegative G-invariant regular Borel measure on X is a scalar multiple of μ. A necessary and sufficient condition for the existence of a nonzero nonnegative G-invariant regular Borel measure on X can be expressed in terms of the modular function of G and the isotropy group of a point of X ([2, Theorem 2.7.11]); such condition is always satisfied for instance if such isotropy group is compact ([2, 2.7.12]).

Example 22.4. If X is a nonempty finite set endowed with the discrete topology then the counting measure is a nonzero nonnegative G-invariant regular Borel measure on X, for any transitive action of a group G on X. If $X = G = \mathbb{R}^n$ and G acts on X by translations then the Lebesgue measure (restricted to the Borel σ-algebra) is a nonzero nonnegative G-invariant regular Borel measure on X. More generally, if a locally compact Hausdorff topological group G acts on itself by left translations then a nonzero nonnegative G-invariant regular Borel measure on G always exists and it is called a *left-invariant Haar measure* on G.

We can now define the abstract generalization of the uniform distribution.

Definition 22.5. Let G and X be as above and assume that a nonzero nonnegative G-invariant regular Borel measure μ on X exists. If M is a Borel subset of X with $0 < \mu(M) < +\infty$ and M is endowed with its Borel σ-algebra $\mathcal{B}$ then the *G-uniform distribution* (or simply *uniform distribution*) on M is the probability measure $\mathbb{P}$ on $\mathcal{B}$ defined by

$$\mathbb{P}(B) = \frac{\mu(B)}{\mu(M)},$$

for all $B \in \mathcal{B}$. An $(M, \mathcal{B})$-valued random object X is said to have a *uniform distribution* if $\mathbb{P}_X$ is a uniform distribution.

23. Quick review of linear and multilinear algebra

This section primarily serves to establish notation and conventions, as prior knowledge of the topics is assumed. All vector spaces considered will be real and finite-dimensional. The bidual space V^{**} of a vector space V will be identified with V itself through the standard natural isomorphism.

If $T : V \to W$ is a linear map, we denote as usual by $T^* : W^* \to V^*$ the *adjoint* (also called *transpose*) of T which is given by $T^*(\alpha) = \alpha \circ T$, for all $\alpha \in W^*$. If V and W are endowed with bases and the dual spaces are endowed with the corresponding dual bases then the matrix that represents T^* is the transpose of the matrix that represents T. Under the natural identification of the bidual spaces with the original spaces, the map T^{**} defined as the transpose of the transpose of T is simply equal to T.

Given vector spaces V and W, the tensor product $V^* \otimes W^*$ of their duals is naturally identified with the space of bilinear forms on $V \times W$ by associating $\alpha \otimes \beta$ to the bilinear form given by

$$(\alpha \otimes \beta)(v, w) = \alpha(v)\beta(w),$$

for all $v \in V$, $w \in W$, $\alpha \in V^*$ and $\beta \in W^*$. In particular, the space $V \otimes W \cong V^{**} \otimes W^{**}$ is naturally identified with the space of bilinear forms on $V^* \times W^*$ by associating $v \otimes w$ to the bilinear form given by

$$(v \otimes w)(\alpha, \beta) = \alpha(v)\beta(w),$$

for all $\alpha \in V^*$, $\beta \in W^*$, $v \in V$ and $w \in W$. These identifications of tensor products with spaces of bilinear forms were already used in Section 11.

We will also often use a natural identification of the space of bilinear forms with a space of linear transformations that is very convenient as it allows us to write several operations involving bilinear forms and linear transformations in terms of compositions and inversions of linear transformations. Since compositions and inversions of linear transformations become products and inversions of matrices once matrix representations are used, the identification of bilinear forms with linear transformations allows for quick translations of the operations with bilinear forms in terms of operations with matrices.

Given vector spaces V, W and a bilinear form $B : V \times W \to \mathbb{R}$, we identify B with the linear transformation from V to W^* given by

$$V \ni v \longmapsto B(v, \cdot) \in W^*, \tag{23.1}$$

where $B(v, \cdot)$ maps $w \in W$ to $B(v, w)$. The map that sends B to (23.1) is a linear isomorphism from the space of bilinear forms on $V \times W$ to the space of linear transformations from V to W^*. The linear transformation (23.1) will be simply denoted by $B : V \to W^*$. If $T : V' \to V$ is a linear map defined in some other vector space V', we can combine B and T to obtain a bilinear form

$$B(T\,\cdot\,,\,\cdot\,) : V' \times W \ni (v', w) \longmapsto B\big(T(v'), w\big) \in \mathbb{R}. \tag{23.2}$$

Under the identification of bilinear forms with linear transformations, the bilinear form (23.2) is identified with the composition $B \circ T : V' \to W^*$. Similarly, if $S : W' \to W$ is a linear map defined in some other vector space W', we can combine B and S to obtain a bilinear form

$$B(\,\cdot\,, S\,\cdot\,) : V \times W' \ni (v, w') \longmapsto B\big(v, S(w')\big) \in \mathbb{R}$$

which is identified with the linear transformation $S^* \circ B : V \to (W')^*$. Hence the bilinear form

$$B(T\,\cdot\,, S\,\cdot\,) : V' \times W' \ni (v', w') \longmapsto B\big(T(v'), S(w')\big) \in \mathbb{R}$$

is identified with the linear transformation $S^* \circ B \circ T$.

Note that the adjoint $B^* : W^{**} \cong W \to V^*$ of the linear transformation $B : V \to W^*$ identified with a bilinear form $B : V \times W \to \mathbb{R}$ is the linear transformation identified with the bilinear form

$$W \times V \ni (w, v) \longmapsto B(v, w) \in \mathbb{R}$$

obtained by switching the variables of B. This implies that if $V = W$ then B is symmetric if and only if $B^* = B$.

A word of caution must be said about matrix representations concerning the identification of bilinear forms with linear transformations discussed above. If $(e_i)_{i=1}^n$ is a basis of V and $(f_i)_{i=1}^m$ is a basis of W, then the matrix that represents the bilinear form $B : V \times W \to \mathbb{R}$ is typically defined as the matrix having $B(e_i, f_j)$ in its i-th row and j-th column. However, if W^* is endowed with the dual basis, the matrix that represents the linear transformation $B : V \to W^*$ that is identified with B is the transpose of the matrix that is normally used to represent the bilinear form B. Thus, if one is going to translate compositions of the form $S^* \circ B \circ T$ into products of matrices, one should be aware that the matrix that must be used for B is the transpose of the usual matrix. This observation is of course irrelevant if $V = W$ and B is symmetric.

23.1. **Correspondence between inner products on V and V*.** Let V be a vector space and $B : V \times V \to \mathbb{R}$ be a symmetric bilinear form. The subspace

$$\big\{v \in V : B(v, w) = 0, \text{ for all } w \in V\big\}$$

of V which is normally known as the kernel of B is simply the kernel $\mathrm{Ker}(B)$ of the linear transformation $B : V \to V^*$ that is identified with B. If B is nondegenerate, i.e., if the kernel of B is zero then $B : V \to V^*$ is a linear isomorphism. Such linear isomorphism can be used to transfer a bilinear form $C : V \times V \to \mathbb{R}$ on V to a bilinear form $C(B^{-1}\cdot, B^{-1}\cdot) : V^* \times V^* \to \mathbb{R}$ on V^*. As discussed above, $C(B^{-1}\cdot, B^{-1}\cdot)$ is identified with the linear transformation from V^* to V given by

$$(B^{-1})^* \circ C \circ B^{-1} = (B^*)^{-1} \circ C \circ B^{-1} = B^{-1} \circ C \circ B^{-1}.$$

Setting $C = B$ we obtain that the bilinear form on V^* that corresponds to B via the isomorphism $B : V \to V^*$ induced by B itself is the bilinear form on V^* that is identified with the linear transformation $B^{-1} : V^* \to V$.

The mapping $B \mapsto B^{-1}$ is a bijection between nondegenerate symmetric bilinear forms on V and nondegenerate symmetric bilinear forms on the dual space V^*. Moreover, B is positive definite if and only if B^{-1} is positive definite and thus the mapping $B \mapsto B^{-1}$ restricts to a bijection between inner products on V and inner products on V^*. This is the standard way[2] of inducing an inner product on the dual space V^* from an inner product on V.

In terms of matrix representations, if V is endowed with some basis and V^* is endowed with the dual basis then the matrix that represents B^{-1} is just the inverse of the matrix that represents B. In particular, if B is an inner product on V then the dual of a B-orthonormal basis of V is a B^{-1}-orthonormal basis of V^* since a basis is orthonormal if and only if the inner product is represented by the identity matrix with respect to such basis. Note also that the isomorphism $B : V \to V^*$ maps a B-orthonormal basis of V to its dual basis. The norms corresponding to the inner products B and B^{-1} are related by the standard formula for the operator norm

$$\|\alpha\| = \sup\big\{|\alpha(v)| : v \in V,\ \|v\| \le 1\big\}, \quad \alpha \in V^*,$$

in which we use the notation $\|\cdot\|$ for both the norm associated to B and for the norm associated to B^{-1}.

Example 23.1. If $B : V \times V \to \mathbb{R}$ and $B' : W \times W \to \mathbb{R}$ are nondegenerate symmetric bilinear forms on vector spaces V and W then, for any linear map $T : V \to W$, there exists a unique linear map $S : W \to V$ that is characterized by the equality

$$B'\big(T(v), w\big) = B\big(v, S(w)\big), \quad v \in V,\ w \in W. \tag{23.3}$$

[2]Physicists call this construction "raising the indexes" of the metric tensor.

The linear map S is usually called the *transpose* of T with respect to B and B'. Identifying B and B' with linear isomorphisms $B : V \to V^*$ and $B' : W \to W^*$, we have that equality (23.3) is equivalent to:

$$B' \circ T = S^* \circ B. \tag{23.4}$$

Taking adjoints on both sides of (23.4) we get an equivalent equality

$$T^* \circ B' = B \circ S,$$

which yields:

$$S = B^{-1} \circ T^* \circ B'.$$

23.2. **A bit of exterior algebra.** Recall that, given a vector space V and a nonnegative integer k, the *k-th exterior power* of V consists of a vector space $\bigwedge_k V$ and an alternating k-linear map

$$V^k \ni (v_1, \ldots, v_k) \longmapsto v_1 \wedge \ldots \wedge v_k \in \bigwedge_k V$$

such that the following property holds: for every vector space W and every alternating k-linear map $B : V^k \to W$ there exists a unique linear transformation $\overline{B} : \bigwedge_k V \to W$ such that $\overline{B}(v_1 \wedge \ldots \wedge v_k) = B(v_1, \ldots, v_k)$, for all $v_1, \ldots, v_k \in V$. In particular, setting $W = \mathbb{R}$ we obtain that the dual space of $\bigwedge_k V$ can be naturally identified with the space of alternating k-linear forms on V by associating $\alpha \in \big(\bigwedge_k V\big)^*$ to the alternating k-linear form:

$$V^k \ni (v_1, \ldots, v_k) \longmapsto \alpha(v_1 \wedge \ldots \wedge v_k) \in \mathbb{R}.$$

If $(e_i)_{i=1}^n$ is a basis of V then

$$e_{i_1} \wedge \ldots \wedge e_{i_k}, \quad 1 \leq i_1 < \cdots < i_k \leq n \tag{23.5}$$

is a basis of $\bigwedge_k V$, so that the dimension of $\bigwedge_k V$ is $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ for $k \leq n$ and zero for $k > n$, where n denotes the dimension of V. In particular, the dimension of $\bigwedge_n V$ is equal to 1.

We have a bilinear map

$$\langle \cdot, \cdot \rangle : \bigwedge_k V \times \bigwedge_k V^* \longrightarrow \mathbb{R}$$

characterized by the equality

$$\langle v_1 \wedge \ldots \wedge v_k, \alpha_1 \wedge \ldots \wedge \alpha_k \rangle = \det\big(\alpha_i(v_j)\big)_{k \times k}, \tag{23.6}$$

for all $v_1, \ldots, v_k \in V$ and $\alpha_1, \ldots, \alpha_k \in V^*$. Such bilinear map induces a linear map

$$\bigwedge_k V^* \ni \omega \longmapsto \langle \cdot, \omega \rangle \in \Big(\bigwedge_k V\Big)^* \tag{23.7}$$

which is easily shown to be an isomorphism using bases of the form (23.5). We will use (23.7) to identify $\bigwedge_k V^*$ with the dual space of $\bigwedge_k V$. This yields

an identification of $\bigwedge_k V^*$ with the space of alternating k-linear forms on V that associates

$$\alpha_1 \wedge \ldots \wedge \alpha_k \in \bigwedge_k V^*$$

to the alternating k-linear form given by

$$(\alpha_1 \wedge \ldots \wedge \alpha_k)(v_1, \ldots, v_k) = \det\big(\alpha_i(v_j)\big)_{k\times k}, \quad v_1, \ldots, v_k \in V,$$

for all $\alpha_1, \ldots, \alpha_k \in V^*$. Similarly, $\bigwedge_k V \cong \bigwedge_k V^{**}$ is identified with the space of alternating k-linear forms on V^* by associating $v_1 \wedge \ldots \wedge v_k \in \bigwedge_k V$ to the alternating k-linear form given by

$$(v_1 \wedge \ldots \wedge v_k)(\alpha_1, \ldots, \alpha_k) = \det\big(\alpha_i(v_j)\big)_{k\times k}, \quad \alpha_1, \ldots, \alpha_k \in V^*$$

for all $v_1, \ldots, v_k \in V$.

Given vector spaces V and W and a linear transformation $T : V \to W$, there exists a unique linear transformation

$$\bigwedge_k T : \bigwedge_k V \longrightarrow \bigwedge_k W$$

satisfying

$$\Big(\bigwedge_k T\Big)(v_1 \wedge \ldots \wedge v_k) = T(v_1) \wedge \ldots \wedge T(v_k)$$

for all $v_1, \ldots, v_k \in V$. We call $\bigwedge_k T$ the *k-th exterior power* of T. For simplicity, we write simply T_* instead of $\bigwedge_k T$ when there is no risk of confusion. Using the identification between the exterior power and the space of alternating multilinear forms, the map T_* is given by

$$(T_*\lambda)(\alpha_1, \ldots, \alpha_k) = \lambda\big(T^*(\alpha_1), \ldots, T^*(\alpha_k)\big),$$

for all $\alpha_1, \ldots, \alpha_k \in W^*$ and every alternating k-linear form λ on V^*. The map

$$\bigwedge_k T^* : \bigwedge_k W^* \longrightarrow \bigwedge_k V^*$$

given by the k-th exterior power of the adjoint of T will be simply denoted by T^* when there is no risk of confusion. Again, using the identification between the exterior power and the space of alternating multilinear forms, we have

$$(T^*\omega)(v_1, \ldots, v_k) = \omega\big(T(v_1), \ldots, T(v_k)\big),$$

for all $v_1, \ldots, v_k \in V$ and every alternating k-linear form ω on W. The map $T^* : \bigwedge_k W^* \to \bigwedge_k V^*$ is usually known as the *pull-back map* of alternating k-linear forms and $T^*\omega$ is the *pull-back* of ω under T.

For $\lambda \in \bigwedge_k V$, $\omega \in \bigwedge_k W^*$ and a linear map $T : V \to W$ we have the equality

$$\langle T_*(\lambda), \omega\rangle = \langle \lambda, T^*(\omega)\rangle$$

which means that $T^* : \bigwedge_k W^* \to \bigwedge_k V^*$ is identified with the adjoint of $T_* : \bigwedge_k V \to \bigwedge_k W$.

If $B : V \times V \to \mathbb{R}$ is a bilinear form, which we identify with a linear map from V to V^*, then

$$\bigwedge_k B : \bigwedge_k V \longrightarrow \bigwedge_k V^* \cong \Big(\bigwedge_k V\Big)^*$$

is identified with a bilinear form on $\bigwedge_k V$ that is characterized by the equality

$$\Big(\bigwedge_k B\Big)(v_1 \wedge \ldots \wedge v_k, w_1 \wedge \ldots \wedge w_k) = \det\big(B(v_i, w_j)\big)_{k\times k},$$

for all $v_1, \ldots, v_k, w_1, \ldots, w_k \in V$. Clearly, if B is symmetric then $\bigwedge_k B$ is symmetric. Moreover, if B is nondegenerate then $B : V \to V^*$ is an isomorphism, so that $\bigwedge_k B : \bigwedge_k V \to \bigwedge_k V^*$ is also an isomorphism and $\bigwedge_k B$ is nondegenerate. Finally, if B is symmetric and positive definite (i.e., B is an inner product) then $\bigwedge_k B$ is also symmetric and positive definite. This can be seen by observing that if $(e_i)_{i=1}^n$ is a B-orthonormal basis for V then (23.5) is a $\bigwedge_k B$-orthonormal basis for $\bigwedge_k V$.

23.3. **Determinant of a linear transformation.** If we identify $\bigwedge_k V^*$ with the space of alternating k-linear forms on V then a basis for the one-dimensional space $\bigwedge_n(\mathbb{R}^n)^*$ is given by the determinant map

$$\det : (\mathbb{R}^n)^n \longrightarrow \mathbb{R}$$

where $\det(v_1, \ldots, v_n)$ is understood as the determinant of the matrix having $v_1, \ldots, v_n \in \mathbb{R}^n$ in its columns (or rows). Moreover, if $(\alpha_i)_{i=1}^n$ is dual to the canonical basis of $\mathbb{R}^n$ then:

$$\det = \alpha_1 \wedge \ldots \wedge \alpha_n.$$

If $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation then the pull-back $T^*\det$ is simply the product of det by the scalar $\det(T)$, which is the determinant of T, i.e., the determinant of the matrix that represents T with respect to the canonical basis. More generally, if V is any n-dimensional vector space and $T : V \to V$ is a linear transformation then $\bigwedge_n V^*$ is a one-dimensional space and thus the linear map $T^* : \bigwedge_n V^* \to \bigwedge_n V^*$ is given by multiplication by a scalar. Such scalar is the determinant of T, i.e., the determinant of the matrix that represents T with respect to an arbitrary basis of V. Here it is crucial that one uses the same basis for the domain and counterdomain of T, otherwise the determinant will depend on the choices of bases.

If V and W are vector spaces having the same dimension n, one cannot in general define a determinant for a linear transformation $T : V \to W$. Namely, in this context it doesn't make sense to have "the same basis" on V and W and $T^* : \bigwedge_n W^* \to \bigwedge_n V^*$ is a linear transformation between distinct one-dimensional spaces and thus it is not multiplication by a scalar. Yet, a determinant becomes well-defined once more structure is added to the vector spaces.

Definition 23.2. If V is a vector space, then a *volume form* on V is a nonzero element ω of $\bigwedge_n V^*$, where n denotes the dimension of V. If (V, ω) is a vector space endowed with a volume form then a *unit volume positive basis* of V is a basis $(e_i)_{i=1}^n$ of V such that $\omega(e_1, \ldots, e_n) = 1$.

Clearly, if ω is a volume form on V then $\{\omega\}$ is a basis of $\bigwedge_n V^*$ and any other volume form on V is of the form $c\,\omega$, for some $c \neq 0$. For every basis $(e_i)_{i=1}^n$ of V there exists a unique volume form on V such that $(e_i)_{i=1}^n$ is a unit volume positive basis of V; such volume form is given by $\alpha_1 \wedge \ldots \wedge \alpha_n$, with $(\alpha_i)_{i=1}^n$ the basis dual to $(e_i)_{i=1}^n$. The volume form on $\mathbb{R}^n$ that makes the canonical basis a unit volume positive basis is the determinant, which is called the *canonical volume form* of $\mathbb{R}^n$.

We can now define determinants for linear transformations between vector spaces of the same dimension endowed with volume forms.

Definition 23.3. Let (V, ω) and (W, ω') be vector spaces having the same dimension and endowed with volume forms. Given a linear transformation $T : V \to W$, its *determinant* is defined as the real number $\det(T)$ such that

$$T^*\omega' = \det(T)\omega,$$

i.e., $\det(T)$ is the unique entry of the 1×1 matrix that represents the linear map $T^* : \bigwedge_n W^* \to \bigwedge_n V^*$ with respect to the bases $\{\omega'\}$ and $\{\omega\}$, where $n = \dim(V) = \dim(W)$.

Note that if $(e_i)_{i=1}^n$ is a unit volume positive basis for V and $(e'_i)_{i=1}^n$ is a unit volume positive basis for W then

$$\det(T) = (T^*\omega')(e_1, \ldots, e_n) = (\alpha'_1 \wedge \ldots \wedge \alpha'_n)\big(T(e_1), \ldots, T(e_n)\big),$$

where $(\alpha'_i)_{i=1}^n$ denotes the basis dual to $(e'_i)_{i=1}^n$. Thus $\det(T)$ is simply the determinant of a matrix that represents T with respect to unit volume positive bases of V and W.

Clearly, if (V, ω), (W, ω') and (Z, ω'') are vector spaces having the same dimension endowed with volume forms and if $T : V \to W$ and $S : W \to Z$ are linear transformations then

$$\det(S \circ T) = \det(S)\det(T).$$

Definition 23.4. If ω is a volume form on a vector space V then the volume form ω^* *induced* on the dual space V^* is defined by requiring that $\{\omega\}$ be the dual basis of $\{\omega^*\}$ when $\bigwedge_n V^*$ is identified with the dual space of $\bigwedge_n V^{**} \cong \bigwedge_n V$ through the isomorphism (23.7) induced by the bilinear pairing (23.6).

Thus, the induced volume form $\omega^* \in \bigwedge_n V$ on the dual space of V is characterized by

$$\langle \omega^*, \omega \rangle = 1,$$

with $\langle \cdot, \cdot \rangle$ defined as in (23.6). It follows that if $(e_i)_{i=1}^n$ is a unit volume positive basis for (V, ω) then the dual basis $(\alpha_i)_{i=1}^n$ is a unit volume positive

basis for (V^*, ω^*), i.e., $\omega^* = e_1 \wedge \ldots \wedge e_n$ if $\omega = \alpha_1 \wedge \ldots \wedge \alpha_n$. Note also that

$$(c\,\omega)^* = \frac{1}{c}\,\omega^*, \tag{23.8}$$

for any $c \neq 0$.

If V and W have the same dimension and are endowed with volume forms and if their dual spaces V^* and W^* are endowed with the induced volume forms then

$$\det(T^*) = \det(T),$$

for any linear transformation $T : V \to W$, where $T^* : W^* \to V^*$ denotes its adjoint. Namely, the pull-back map $T^* : \bigwedge_n W^* \to \bigwedge_n V^*$ associated to T is identified with the adjoint of the pull-back map $(T^*)^* = T_* : \bigwedge_n V \to \bigwedge_n W$ associated to T^*.

23.4. **Volume forms and Lebesgue measure.** Unlike the space $\mathbb{R}^n$, an abstract n-dimensional vector space V does not have a canonical translation invariant locally finite measure. If we identify V with $\mathbb{R}^n$ using a linear isomorphism then such identification can be used to carry the Lebesgue measure $\mathfrak{m}$ of $\mathbb{R}^n$ to V, but the measure obtained on V will depend on the chosen isomorphism. Two measures obtained by this procedure will be a scalar multiple of each other (which follows from (23.10) below) and we will call anyone of them a *Lebesgue measure* on V. Here we will show that a specific Lebesgue measure on V is determined by the choice of a volume form on V, which justifies the name "volume form".

We start by recalling that if $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation then

$$\mathfrak{m}\big(T[B]\big) = |\det(T)|\,\mathfrak{m}(B), \tag{23.9}$$

for any Lebesgue measurable subset B of $\mathbb{R}^n$. Equality (23.9) can be obtained from the change of variables theorem for Lebesgue integration (Theorem 20.1), though in fact it is often proven as a lemma which is used in the proof of Theorem 20.1. If T is a linear isomorphism, equality (23.9) is equivalent to:

$$T_*\mathfrak{m} = \frac{1}{|\det(T)|}\,\mathfrak{m}. \tag{23.10}$$

Since a linear isomorphism of $\mathbb{R}^n$ maps Lebesgue measurable sets to Lebesgue measurable sets, we can define the *Lebesgue σ-algebra* of an arbitrary n-dimensional vector space V as the σ-algebra consisting of the images of the Lebesgue measurable subsets of $\mathbb{R}^n$ under any linear isomorphism from $\mathbb{R}^n$ to V. We are now ready to define the Lebesgue measure associated to a volume form.

Definition 23.5. Let V be a vector space and ω be a volume form on V. The *Lebesgue measure* on V associated to ω is the measure $\mathfrak{m}_\omega$ on the Lebesgue σ-algebra of V defined by $T_*\mathfrak{m}$, where $\mathfrak{m}$ denotes the Lebesgue measure of $\mathbb{R}^n$ and $T : \mathbb{R}^n \to V$ is any linear isomorphism such that $T^*\omega = \det$

(equivalently, $T : \mathbb{R}^n \to V$ is any linear isomorphism that maps the canonical basis of $\mathbb{R}^n$ to a unit volume positive basis of (V, ω)).

Using equality (23.10) we show that the Lebesgue measure $\mathfrak{m}_\omega$ is indeed well-defined, i.e., it does not depend on the choice of the linear isomorphism T. Namely, if $S : \mathbb{R}^n \to V$ is another linear isomorphism with $S^*\omega = \det$ then $S = T \circ L$, with $L : \mathbb{R}^n \to \mathbb{R}^n$ a linear isomorphism with $L^* \det = \det$, i.e., $\det(L) = 1$. We then have $L_*\mathfrak{m} = \mathfrak{m}$ and thus $T_*\mathfrak{m} = S_*\mathfrak{m}$.

Clearly, the Lebesgue σ-algebra of V is the completion of the Borel σ-algebra of V with respect to any Lebesgue measure $\mathfrak{m}_\omega$ (as it is the case in $\mathbb{R}^n$). Note also that the Lebesgue measure of $\mathbb{R}^n$ is the Lebesgue measure $\mathfrak{m}_{\det}$ associated to the canonical volume form of $\mathbb{R}^n$.

The fact that the Lebesgue measure of $\mathbb{R}^n$ is invariant under translations yields that any Lebesgue measure $\mathfrak{m}_\omega$ on a vector space V is also invariant under translations, as shown below.

Proposition 23.6. *If V is a vector space endowed with a volume form ω then the Lebesgue measure $\mathfrak{m}_\omega$ is invariant under translations, i.e.,*

$$\mathfrak{m}_\omega(B + v) = \mathfrak{m}_\omega(B),$$

for any $v \in V$ and any Lebesgue measurable subset B of V, where:

$$B + v = \big\{x + v : x \in B\big\}.$$

Proof. If $T : \mathbb{R}^n \to V$ is a linear isomorphism with $T^*\omega = \det$ then:

$$\mathfrak{m}_\omega(B+v) = \mathfrak{m}\big(T^{-1}[B+v]\big) = \mathfrak{m}\big(T^{-1}[B]+T^{-1}(v)\big) = \mathfrak{m}\big(T^{-1}[B]\big) = \mathfrak{m}_\omega[B],$$

since the Lebesgue measure $\mathfrak{m}$ of $\mathbb{R}^n$ is invariant under translations. □

We can now generalize (23.9) to a linear map between arbitrary vector spaces having the same dimension and endowed with volume forms.

Proposition 23.7. *Let (V, ω) and (W, ω') be vector spaces having the same dimension endowed with volume forms ω and ω'. If $T : V \to W$ is a linear transformation, then*

$$\mathfrak{m}_{\omega'}\big(T[B]\big) = |\det(T)|\, \mathfrak{m}_\omega(B),$$

for any Lebesgue measurable subset B of V. In particular, if T is an isomorphism then:

$$T_*\mathfrak{m}_\omega = \frac{1}{|\det(T)|}\, \mathfrak{m}_{\omega'}.$$

Proof. Let $L_1 : \mathbb{R}^n \to V$, $L_2 : \mathbb{R}^n \to W$ be linear isomorphisms with

$$L_1^*\, \omega = \det, \quad L_2^*\, \omega' = \det,$$

so that

$$\mathfrak{m}_\omega = (L_1)_*\mathfrak{m}, \quad \mathfrak{m}_{\omega'} = (L_2)_*\mathfrak{m}$$

and $\det(L_1) = \det(L_2) = 1$ if $\mathbb{R}^n$ is endowed with its canonical volume form det. Setting $T' = L_2^{-1} \circ T \circ L_1 : \mathbb{R}^n \to \mathbb{R}^n$ and $B' = L_1^{-1}[B]$ then $L_2^{-1}\big[T[B]\big] = T'[B']$ and

$$\mathfrak{m}_{\omega'}\big(T[B]\big) = \mathfrak{m}\big(T'[B']\big) = |\det(T')|\,\mathfrak{m}(B') = |\det(T')|\,\mathfrak{m}_{\omega}(B),$$

where $\det(T') = \det(L_2)^{-1}\det(T)\det(L_1) = \det(T)$. □

Corollary 23.8. *If V is a vector space endowed with a volume form ω then*

$$\mathfrak{m}_{c\,\omega} = |c|\,\mathfrak{m}_{\omega},$$

for any $c \neq 0$.

Proof. Use Proposition 23.7 with $W = V$, $\omega' = c\,\omega$ and T the identity map of V. □

Corollary 23.9. *If V is a vector space and ω and ω' are volume forms on V then $\mathfrak{m}_{\omega} = \mathfrak{m}_{\omega'}$ if and only if $\omega = \omega'$ or $\omega = -\omega'$.* □

Example 23.10. Let $B : V \times V \to \mathbb{R}$ be a symmetric bilinear form. If ω is a volume form on V and ω^* is the volume form induced on V^*, we can consider the determinant $\det(B)$ of the linear transformation $B : V \to V^*$ that is identified with B with respect to ω and ω^*. If ω is replaced with $c\,\omega$ for some $c \neq 0$ then ω^* is replaced with $\frac{1}{c}\,\omega^*$ (see (23.8)) and therefore the determinant of B gets multiplied by $\frac{1}{c^2}$. If B is nondegenerate, this implies that there exists a volume form ω_B on V which makes the absolute value of the determinant of B with respect to ω_B and ω_B^* equal to 1; moreover, the volume form ω_B is unique up to a sign. More explicitly, we have

$$\omega_B = \pm\sqrt{|\det(B)|}\,\omega,$$

where ω is an arbitrary volume form on V and $\det(B)$ is the determinant of B with respect to ω and ω^*. Given a basis $(e_i)_{i=1}^n$ of V with dual basis $(\alpha_i)_{i=1}^n$, by setting $\omega = \alpha_1 \wedge \ldots \wedge \alpha_n$ we obtain

$$\omega_B = \pm\sqrt{|\det(B)|}\;\alpha_1 \wedge \ldots \wedge \alpha_n,$$

where $\det(B)$ is the determinant of B with respect to ω and ω^*, i.e., the determinant of the matrix that represents the bilinear form B with respect to the basis $(e_i)_{i=1}^n$. In particular, if the basis $(e_i)_{i=1}^n$ is B-orthogonal and $|B(e_i, e_i)| = 1$ for all $i = 1, \ldots, n$ then $\omega_B = \pm\alpha_1 \wedge \ldots \wedge \alpha_n$. Note that if B is positive definite (i.e., B is an inner product) then $\det(B)$ is actually positive for any choice of ω, as the sign of $\det(B)$ is independent of the volume form ω and $\det(B) = 1$ if $\omega = \alpha_1 \wedge \ldots \wedge \alpha_n$ with $(\alpha_i)_{i=1}^n$ dual to a B-orthonormal basis $(e_i)_{i=1}^n$ of V.

In order to fix a sign for ω_B one could choose an orientation of V and demand that ω_B take a positive value on positive bases of V. However, the Lebesgue measure associated to ω_B does not depend on the sign of ω_B and therefore we have a Lebesgue measure $\mathfrak{m}_B$ associated to any nondegenerate symmetric bilinear form $B : V \times V \to \mathbb{R}$. In particular, there is a Lebesgue

measure associated to any choice of inner product. For example, in $\mathbb{R}^n$ the usual Lebesgue measure is associated to the canonical inner product.

24. The multivariate normal distribution

After the conclusion of our review of linear and multilinear algebra we are ready to generalize the normal distribution to vector spaces. A convenient way to define normal distributions on vector spaces is by relating them to normal distributions on the real line by means of linear functionals. A normal distribution on a vector space is also known as a *Gaussian measure*. Though the theory of Gaussian measures on infinite-dimensional topological vector spaces has been widely studied, we will here focus only on the much simpler finite-dimensional case.

Definition 24.1. Let V be a real finite-dimensional vector space. A V-valued random vector X is said to have a *normal distribution* (alternatively, X is a *normal random vector* or $\mathbb{P}_X$ is a *normal distribution* on V) if the random variable $\alpha(X)$ has a normal distribution for all $\alpha \in V^*$.

Since the product of a normal random variable by a real number is normal (Proposition 21.3) it follows that a normal random variable is the same thing as a normal $\mathbb{R}$-valued random vector in the sense of Definition 24.1. Moreover, since a normal random variable always has a finite variance, we have that a normal random vector is always square integrable and therefore it has a well-defined expected value $E(X) \in V$ and a well-defined variance $\mathrm{Var}(X) \in V \otimes V$.

As in the univariate case, the expected value and the variance completely determine a normal distribution.

Proposition 24.2. *If V is a real finite-dimensional vector space then two V-valued normal random vectors with the same expected value and the same variance have the same distribution.*

Proof. Follows directly from Propositions 14.2 and 21.5. □

Proposition 24.2 motivates the notation introduced below.

Definition 24.3. Given a real finite-dimensional vector space V, a vector $\mu \in V$ and a positive semi-definite symmetric bilinear form $\Sigma \in V \otimes V$ on V^*, we write

$$X \sim N(\mu, \Sigma) \tag{24.1}$$

if X is a V-valued random vector with $E(X) = \mu$ and $\mathrm{Var}(X) = \Sigma$. In this case we also say that the probability measure $\mathbb{P}_X$ is a *normal distribution with mean μ and variance Σ*.

We will prove later in this section (Proposition 24.11) that for every $\mu \in V$ and every positive semi-definite symmetric bilinear form $\Sigma \in V \otimes V$ there exists a V-valued random vector X with $X \sim N(\mu, \Sigma)$. It will then follow that the normal distribution on V can be thought as a family of probability

measures on the Borel σ-algebra of V indexed (in a one-to-one manner) by the parameters $\mu \in V$ and $\Sigma \in V \otimes V$, with Σ positive semi-definite and symmetric. As in the univariate case, it is more common here to call μ the mean instead of expected value. Note that Σ corresponds to what we denoted by σ^2 in the univariate case, not to σ.

The following result is a trivial consequence of Definition 24.1 and of the fact that the sum of a normal random variable with a constant is again normal (Proposition 21.3).

Proposition 24.4. *Let V and W be real finite-dimensional vector spaces and X be a V-valued normal random vector. If $T : V \to W$ is a linear map and $w \in W$ then $T(X) + w$ is a W-valued normal random vector.* □

Remark 24.5. If V is a real finite-dimensional vector space and X is a normal V-valued random vector such that $\mathbb{P}(X \in W) = 1$ for some subspace W of V then we can regard X as a W-valued random vector, possibly by modifying X on a set of probability zero. Clearly, X remains normal when regarded as a W-valued random vector as every linear functional on W admits a linear extension to V.

It would be nice to have a concrete description of the distribution of a normal V-valued random vector X in terms of a probability density function with respect to a Lebesgue measure, like in the univariate case. However, if the variance $\mathrm{Var}(X)$ is degenerate then the support of $\mathbb{P}_X$ is contained in a proper affine subspace of V (Corollary 11.6). Since a proper affine subspace has null Lebesgue measure, in this case $\mathbb{P}_X$ will not be absolutely continuous with respect to a Lebesgue measure. We will then work first with the case in which X is *nondegenerate*, meaning that the variance $\mathrm{Var}(X)$ is nondegenerate (and thus positive definite).

A good place to begin searching for probability density functions is with exponentials of (multivariate) polynomials of degree two. In coordinate-free language, a (real-valued) polynomial with degree less than or equal to 2 on a real finite-dimensional vector space V is a map $p : V \to \mathbb{R}$ of the form

$$p(x) = B(x, x) + \gamma(x) + c, \quad x \in V,$$

with $B : V \times V \to \mathbb{R}$ a symmetric bilinear form, $\gamma \in V^*$ and $c \in \mathbb{R}$. The coefficients B, γ and c of the polynomial p are determined from the map p, which can be seen by noting that, for any $x \in V$, the coefficients of the polynomial $\mathbb{R} \ni t \mapsto p(tx) \in \mathbb{R}$ are $B(x, x)$, $\gamma(x)$ and c. Using a B-orthogonal basis of V one can easily show (as in the proof of Lemma 24.6 below) that the exponential of p will have an infinite integral with respect to a Lebesgue measure on V unless B is negative definite. Therefore we only consider exponentials of second degree polynomials with a negative definite leading coefficient.

Lemma 24.6. *Let V be a real finite-dimensional vector space, ω be a volume form on V and denote by $\mathfrak{m}_\omega$ the corresponding Lebesgue measure on V. If*

$B : V \times V \to \mathbb{R}$ is a positive definite symmetric bilinear form then the integral

$$\int_V e^{-B(x,x)}\,\mathrm{d}\mathfrak{m}_\omega(x)$$

is a (finite) positive number and therefore

$$f_X(x) = c\,e^{-B(x,x)}, \quad x \in V$$

is the probability density function of some V-valued random vector X with respect to $\mathfrak{m}_\omega$ for a unique $c > 0$. Moreover, a V-valued random vector X with probability density function f_X with respect to $\mathfrak{m}_\omega$ is normal.

Proof. Let $T : V \to \mathbb{R}^n$ be a linear map that sends a B-orthonormal basis of V to the canonical basis of $\mathbb{R}^n$, so that

$$\langle T(x), T(y)\rangle = B(x, y), \tag{24.2}$$

for all $x, y \in V$, where $\langle\cdot,\cdot\rangle$ denotes the canonical inner product of $\mathbb{R}^n$. By Proposition 23.7 we have

$$T_*\mathfrak{m}_\omega = \frac{1}{|\det(T)|}\,\mathfrak{m}, \tag{24.3}$$

where $\mathfrak{m}$ denotes the Lebesgue measure of $\mathbb{R}^n$ and $\det(T)$ is the determinant of T with respect to the volume forms ω and det.

Denoting by $f : V \to \mathbb{R}$ the map given by $f(x) = e^{-B(x,x)}$, for all x in V, the abstract "change of variables" theorem for push-forward measures (Proposition 9.1) gives:

$$\int_V f\,\mathrm{d}\mathfrak{m}_\omega = \frac{1}{|\det(T)|}\int_{\mathbb{R}^n} f \circ T^{-1}\,\mathrm{d}\mathfrak{m}.$$

By (24.2), for all $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ we have

$$(f \circ T^{-1})(z) = e^{-\langle z,z\rangle} = \prod_{i=1}^{n} e^{-z_i^2}$$

and then an application of Fubini–Tonelli's Theorem yields that the integral of $f \circ T^{-1}$ with respect to $\mathfrak{m}$ is finite. This concludes the proof of the first part of the lemma.

Now let $c > 0$ be such that $f_X = cf$ has integral with respect to $\mathfrak{m}_\omega$ equal to 1 and let us prove that a V-valued random vector X having f_X as a probability density function with respect to $\mathfrak{m}_\omega$ is normal. To this aim, it suffices to prove that $\alpha(X)$ is normal for all $\alpha \in V^*$ with $B^{-1}(\alpha, \alpha) = 1$, as the product of a normal random variable by a real number is normal. We can then assume that $\alpha = \alpha_1$, with $(\alpha_i)_{i=1}^n$ a B^{-1}-orthonormal basis of V^*.

Let $(e_i)_{i=1}^n$ be the basis of V given by $e_i = B^{-1}(\alpha_i)$, $i = 1, \ldots, n$, so that $(e_i)_{i=1}^n$ is B-orthonormal, $(\alpha_i)_{i=1}^n$ is dual to $(e_i)_{i=1}^n$ and the linear map $T : V \to \mathbb{R}^n$ that sends $(e_i)_{i=1}^n$ to the canonical basis of $\mathbb{R}^n$ has $(\alpha_i)_{i=1}^n$ as its coordinate functionals. Using (24.3) and the results discussed in Example 8.7

we get that a probability density function for $T(X)$ with respect to $\mathfrak{m}$ is given by

$$f_{T(X)} = \frac{1}{|\det(T)|}\, f_X \circ T^{-1} = \frac{c}{|\det(T)|}\, f \circ T^{-1},$$

so that

$$f_{T(X)}(z) = \frac{c}{|\det(T)|} \prod_{i=1}^{n} e^{-z_i^2},$$

for all $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$. It now follows from Proposition 18.10 that a probability density function for $\alpha(X) = \alpha_1(X)$ is given by

$$f_{\alpha(X)}(z_1) = c_1 e^{-z_1^2}, \quad z_1 \in \mathbb{R},$$

for some positive constant c_1 and hence $\alpha(X)$ is normal. □

Let $Z_1, \ldots, Z_n$ be independent standard normal random variables and consider the $\mathbb{R}^n$-valued random vector $Z = (Z_1, \ldots, Z_n)$. From (21.2) and Proposition 18.10 we see that a probability density function $f_Z : \mathbb{R}^n \to \mathbb{R}$ for Z with respect to the Lebesgue measure $\mathfrak{m}$ is given by

$$f_Z(z) = \frac{1}{(2\pi)^{\frac{n}{2}}}\, e^{-\frac{1}{2}\langle z, z\rangle}, \tag{24.4}$$

for all $z \in \mathbb{R}^n$, where $\langle \cdot, \cdot \rangle$ denotes the canonical inner product of $\mathbb{R}^n$. Applying Lemma 24.6 with $V = \mathbb{R}^n$, $\omega = \det$ and $B(z, z) = \frac{1}{2}\langle z, z\rangle$ we obtain that Z is normal. Clearly $E(Z) = 0$ and the covariance matrix of Z is the identity matrix, as the covariance of independent random variables is zero (Corollary 15.11). Thus, $\mathrm{Var}(Z) \in \mathbb{R}^n \otimes \mathbb{R}^n$ is the canonical inner product of $\mathbb{R}^{n*}$ (i.e., the inner product that makes the dual of the canonical basis of $\mathbb{R}^n$ orthonormal).

Definition 24.7. An $\mathbb{R}^n$-valued random vector Z is said to have a *standard normal distribution* (alternatively, Z is a *standard normal random vector* or $\mathbb{P}_Z$ is the *standard normal distribution* on $\mathbb{R}^n$) if (24.4) is a probability density function for Z with respect to Lebesgue measure.

We can now use (24.4) to obtain an explicit formula for a probability density function for any nondegenerate normal random vector. Let then V be a real-finite dimensional vector space, $\mu \in V$ and $\Sigma \in V \otimes V$ be a positive definite symmetric bilinear form on V^*. If $T : \mathbb{R}^n \to V$ is a linear isomorphism that maps the canonical basis of $\mathbb{R}^n$ to a Σ^{-1}-orthonormal basis of V then

$$\Sigma^{-1}\big(T(z), T(w)\big) = \langle z, w\rangle, \tag{24.5}$$

for all $z, w \in \mathbb{R}^n$ and

$$\langle T^*(\alpha), T^*(\beta)\rangle = \Sigma(\alpha, \beta), \tag{24.6}$$

for all $\alpha, \beta \in V^*$, where $\langle \cdot, \cdot \rangle$ denotes both the canonical inner product of $\mathbb{R}^n$ and of $\mathbb{R}^{n*}$. Equality (24.6) says that $T \otimes T : \mathbb{R}^n \otimes \mathbb{R}^n \to V \otimes V$ maps

the canonical inner product of $\mathbb{R}^{n*}$ to Σ and therefore, if Z is a standard normal $\mathbb{R}^n$-valued random vector we have (Corollary 11.9):

$$\mathrm{Var}\big(T(Z)\big) = \Sigma.$$

Setting $X = T(Z) + \mu$, we then conclude using Proposition 24.4 that X is a V-valued normal random vector with $E(X) = \mu$ and $\mathrm{Var}(X) = \Sigma$.

Let us now write down a explicit formula for a probability density function for X. Let ω be a volume form on V, ω^* be the induced volume form on the dual space V^* and $\mathfrak{m}_\omega$ denote the Lebesgue measure on V associated to ω. By Proposition 23.7 we have

$$T_*\mathfrak{m} = \frac{1}{|\det(T)|}\,\mathfrak{m}_\omega, \tag{24.7}$$

where $\mathfrak{m}$ denotes the Lebesgue measure on $\mathbb{R}^n$ and $\det(T)$ denotes the determinant of T with respect to the volume forms det and ω. Using (24.7), the translation invariance of $\mathfrak{m}_\omega$ and the results discussed in Example 8.7 we obtain that a probability density function for $X = T(Z) + \mu$ with respect to $\mathfrak{m}_\omega$ is given by

$$f_X(x) = \frac{1}{|\det(T)|}\, f_Z\big(T^{-1}(x-\mu)\big),$$

for all $x \in V$. Now (24.4) and (24.5) yield:

$$f_X(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\det(T)|}\, e^{-\frac{1}{2}\Sigma^{-1}(x-\mu,x-\mu)}, \quad x \in V. \tag{24.8}$$

The formula for f_X should involve only μ and Σ, so we must get rid of $\det(T)$ somehow. To this aim, note that equality (24.6) can be rewritten as

$$T \circ \eta \circ T^* = \Sigma, \tag{24.9}$$

where $\eta : \mathbb{R}^{n*} \to \mathbb{R}^n$ is the linear map identified with the canonical inner product of $\mathbb{R}^{n*}$. Now endowing $\mathbb{R}^{n*}$ with the volume form induced on the dual space by det we have $\det(\eta) = 1$ and therefore taking determinants on both sides of the equality (24.9) we obtain:

$$\big(\det(T)\big)^2 = \det(\Sigma). \tag{24.10}$$

Taking (24.8) and (24.10) together and keeping Proposition 24.2 in mind we establish the following result.

Proposition 24.8. *Let V be a real finite-dimensional vector space, $\mu \in V$ and $\Sigma \in V \otimes V$ be a positive definite symmetric bilinear form on V^*. A V-valued random vector X satisfying $X \sim N(\mu, \Sigma)$ exists. Moreover, if ω is a volume form on V then a V-valued random vector X satisfies $X \sim N(\mu, \Sigma)$ if and only if the map $f_X : V \to \mathbb{R}$ defined by*

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}}\, e^{-\frac{1}{2}\Sigma^{-1}(x-\mu,x-\mu)}, \quad x \in V \tag{24.11}$$

is a probability density function for X with respect to the Lebesgue measure $\mathfrak{m}_\omega$ on V associated to ω, where $\det(\Sigma)$ denotes the determinant of the linear

map $\Sigma : V^* \to V$ with respect to the volume forms ω^* and ω (as usual, ω^* denotes the volume form on the dual space V^* induced by ω). □

Corollary 24.9. *Let V be a real finite-dimensional vector space and X be a V-valued random vector. We have that X is normal with nondegenerate variance if and only if X admits a probability density function with respect to a Lebesgue measure on V that is equal to the exponential of a second degree polynomial $p : V \to \mathbb{R}$ with a negative definite leading coefficient.*

Proof. If $X \sim N(\mu, \Sigma)$ with $\mu \in V$ and nondegenerate $\Sigma \in V \otimes V$ then f_X defined by (24.11) is a probability density function for X with respect to a Lebesgue measure $\mathfrak{m}_\omega$ and f_X is the exponential of a second degree polynomial whose leading coefficient $-\frac{1}{2}\Sigma^{-1}$ is negative definite. Conversely, let

$$p(x) = -B(x, x) + \gamma(x) + c, \quad x \in V$$

be a second degree polynomial with $B : V \times V \to \mathbb{R}$ bilinear symmetric positive definite, $\gamma \in V^*$ and $c \in \mathbb{R}$. Assume that

$$V \ni x \longmapsto e^{p(x)} \in \mathbb{R} \tag{24.12}$$

is a probability density function for X with respect to a Lebesgue measure $\mathfrak{m}_\omega$, where ω is a volume form on V. Setting

$$\Sigma = \frac{1}{2}\, B^{-1} \in V \otimes V, \quad \mu = \Sigma(\gamma) = \frac{1}{2}\, B^{-1}(\gamma) \in V$$

we obtain that p and the polynomial in the exponent in (24.11) are equal up to the independent term. As both (24.12) and (24.11) have integral equal to 1 with respect to $\mathfrak{m}_\omega$, it must be the case that the maps (24.12) and (24.11) are equal and hence $X \sim N(\mu, \Sigma)$. □

Remark 24.10. Recall from Proposition 11.5 that if a V-valued square integrable random vector X has a degenerate variance $\mathrm{Var}(X)$ then the support of $\mathbb{P}_X$ is contained in a proper affine subspace of V which is a translation of the vector subspace W of V annihilated by the kernel of the positive semi-definite symmetric bilinear form $\mathrm{Var}(X) : V^* \times V^* \to \mathbb{R}$. Moreover, such translation of W is the smallest affine subspace of V containing the support of $\mathbb{P}_X$. Since an affine subspace of V containing the support of $\mathbb{P}_X$ must contain the expected value $E(X)$ (Remark 11.2) we have that if $E(X) = 0$ then the support of $\mathbb{P}_X$ is actually contained in the vector subspace W. When X is regarded as a W-valued random vector (possibly by modifying X on a set of probability zero), the variance of X becomes nondegenerate as the support of the distribution of X is not contained in a proper affine subspace of W.

The remark above gives us the hint of how to construct a normal V-valued random vector with prescribed degenerate variance.

Proposition 24.11. *Let V be a real finite-dimensional vector space, $\mu \in V$ and $\Sigma \in V \otimes V$ be a positive semi-definite symmetric bilinear form on V^*. There exists a V-valued random vector X with $X \sim N(\mu, \Sigma)$.*

Proof. We can assume without loss of generality that $\mu = 0$, as the general case is obtained from this particular case by replacing X with $X + \mu$. The nondegenerate case was already handled in Proposition 24.8. If Σ is degenerate, let W be the subspace of V annihilated by the kernel of Σ, so that $\mathrm{Ker}(\Sigma) = W^{\mathrm{o}}$ is the annihilator of W. If $i : W \to V$ denotes the inclusion map and Y is a normal W-valued random vector with $E(Y) = 0$ then $X = i(Y)$ is a normal V-valued random vector with $E(X) = 0$ and:

$$\mathrm{Var}(X) = (i \otimes i)\big(\mathrm{Var}(Y)\big).$$

More concretely, the map $i \otimes i$ is given by

$$(i \otimes i)(B)(\alpha, \beta) = B(\alpha|_W, \beta|_W),$$

for all $\alpha, \beta \in V^*$ and any bilinear form $B : W^* \times W^* \to \mathbb{R}$. Since the restriction map

$$V^* \ni \alpha \longmapsto \alpha|_W \in W^* \tag{24.13}$$

is a surjective linear map whose kernel is $W^{\mathrm{o}} = \mathrm{Ker}(\Sigma)$, we have that the positive semi-definite symmetric bilinear form $\Sigma : V^* \times V^* \to \mathbb{R}$ passes to the quotient through (24.13) and defines a positive definite symmetric bilinear form $\Sigma_1 : W^* \times W^* \to \mathbb{R}$ such that

$$\Sigma_1(\alpha|_W, \beta|_W) = \Sigma(\alpha, \beta),$$

for all $\alpha, \beta \in V^*$. The latter equality means that $(i \otimes i)(\Sigma_1) = \Sigma$ and the conclusion is now obtained by using Proposition 24.8 to get $Y \sim N(0, \Sigma_1)$ and by setting $X = i(Y)$. □

We need a lemma for the proof of our next result.

Lemma 24.12. *If $(X_i)_{i=1}^n$ is a finite independent family of normal random variables then any linear combination $\sum_{i=1}^n c_i X_i$ is a normal random variable, where $c_1, \ldots, c_n \in \mathbb{R}$.*

Proof. We can assume without loss of generality that all X_i are nondegenerate, as a degenerate normal random variable is almost surely constant and the sum of a normal random variable with a random variable that is almost surely constant is normal. If $X_i \sim N(\mu_i, \sigma_i^2)$ with $\mu_i \in \mathbb{R}$, $\sigma_i > 0$, $i = 1, \ldots, n$, then Propositions 21.5 and 18.10 yield a probability density function

$$f_X(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma}\, e^{p(x)}, \quad x \in \mathbb{R}^n$$

for the $\mathbb{R}^n$-valued random vector $X = (X_1, \ldots, X_n)$ with respect to Lebesgue measure, where

$$p(x) = -\sum_{i=1}^n \frac{(x_i - \mu_i)^2}{2\sigma_i^2}, \quad x \in \mathbb{R}^n$$

and $\sigma = \sigma_1 \cdots \sigma_n$. Since p is a second degree polynomial with a negative definite leading coefficient, Corollary 24.9 implies that the random vector X

is normal. The conclusion now follows by noting that $\sum_{i=1}^n c_i X_i = \alpha(X)$ for a linear functional $\alpha \in \mathbb{R}^{n*}$. □

Given a finite family $(V_i)_{i\in I}$ of vector spaces, we denote by $\bigoplus_{i\in I} V_i$ its *external direct sum*, which is the cartesian product $\prod_{i\in I} V_i$ endowed with the operations defined coordinatewise. Let us show now that by combining independent V_i-valued normal random vectors into a $\bigoplus_{i\in I} V_i$-valued random vector we obtain again a normal random vector.

Proposition 24.13. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and for $i = 1, \dots, n$, let V_i be a real finite-dimensional vector space and $X_i : \Omega \to V_i$ be a normal random vector. If the family $(X_i)_{i=1}^n$ is independent and $V = \bigoplus_{i=1}^n V_i$ then the random vector $X = (X_i)_{i=1}^n : \Omega \to V$ is normal.*

Proof. We have to check that $\alpha(X)$ is normal for any $\alpha \in V^*$. A linear functional $\alpha \in V^*$ is of the form

$$\alpha(x_1, \dots, x_n) = \sum_{i=1}^n \alpha_i(x_i), \quad x_1 \in V_1, \ \dots, \ x_n \in V_n, \tag{24.14}$$

with $\alpha_i \in V_i^*$ for $i = 1, \dots, n$. We have $\alpha(X) = \sum_{i=1}^n \alpha_i(X_i)$ and the conclusion follows from Lemma 24.12 keeping in mind that $\big(\alpha_i(X_i)\big)_{i=1}^n$ is an independent family of normal random variables (Proposition 18.6). □

Recall that two independent (square integrable) random variables have zero covariance (Corollary 15.11) and that, in general, the condition of having zero covariance is much weaker than independence (Example 19.3). It turns out that for a finite family of random variables that constitute the coordinates of a normal random vector the condition of pairwise zero covariance is equivalent to independency.

Proposition 24.14. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and for $i = 1, \dots, n$, let V_i be a real finite-dimensional vector space and $X_i : \Omega \to V_i$ be a random vector. If the random vector $X = (X_i)_{i=1}^n : \Omega \to \bigoplus_{i=1}^n V_i$ is normal and if $\operatorname{Cov}(X_i, X_j) = 0$ for all $i, j = 1, \dots, n$ with $i \neq j$ then the family $(X_i)_{i=1}^n$ is independent.*

Proof. By replacing X_i with $X_i - E(X_i)$ we can assume without loss of generality that $E(X_i) = 0$, for all $i = 1, \dots, n$ (keep in mind Proposition 18.6). Moreover, replacing V_i with the smallest subspace of V_i containing the support of $\mathbb{P}_{X_i}$ (which is the subspace annihilated by the kernel of $\operatorname{Var}(X_i)$) we can assume without loss of generality that $\operatorname{Var}(X_i)$ is nondegenerate, for all $i = 1, \dots, n$ (see Remarks 24.5 and 24.10).

The dual space of $V = \bigoplus_{i=1}^n V_i$ can be identified with $\bigoplus_{i=1}^n V_i^*$ by associating $\alpha \in V^*$ with the sequence $(\alpha_i)_{i=1}^n \in \bigoplus_{i=1}^n V_i^*$ such that (24.14) holds. The fact that $\operatorname{Cov}(X_i, X_j) = 0$ for all $i \neq j$ implies that the variance $\Sigma = \operatorname{Var}(X) : V^* \times V^* \to \mathbb{R}$ of X is given by

$$\Sigma\big((\alpha_1, \dots, \alpha_n), (\beta_1, \dots, \beta_n)\big) = \sum_{i=1}^n \Sigma_i(\alpha_i, \beta_i),$$

for all $\alpha_i, \beta_i \in V_i^*$, $i = 1, \dots, n$, where $\Sigma_i = \mathrm{Var}(X_i)$ for $i = 1, \dots, n$. Since each Σ_i is nondegenerate we have that also Σ is nondegenerate and thus Proposition 24.8 yields that a probability density function of X with respect to a Lebesgue measure of V is of the form

$$f_X(x) = c \prod_{i=1}^{n} e^{-\frac{1}{2}\Sigma_i^{-1}(x_i, x_i)}, \quad x = (x_1, \dots, x_n) \in V,$$

for some $c > 0$. Since the product of Lebesgue measures on the spaces V_i is a Lebesgue measure on V, Proposition 18.10 applies and allows us to conclude that the family $(X_i)_{i=1}^n$ is independent. □

Corollary 24.15. *Let V be a real finite-dimensional vector space and let $T_i : V \to W_i$, $i = 1, \dots, n$, be linear maps, where each W_i is a real finite-dimensional vector space. If X is a normal V-valued random vector and if $\mathrm{Cov}\big(T_i(X), T_j(X)\big) = 0$ for all $i, j = 1, \dots, n$ with $i \neq j$ then the family $\big(T_i(X)\big)_{i=1}^n$ is independent.*

Proof. Set $W = \bigoplus_{i=1}^n W_i$ and let $T : V \to W$ be the linear map whose i-th coordinate is T_i, for all $i = 1, \dots, n$. The conclusion follows from Proposition 24.14 noting that $T(X)$ is normal. □

Corollary 24.16. *Let V be a real finite-dimensional vector space and X be a V-valued normal random vector with nondegenerate variance $\Sigma \in V \otimes V$. If $V = \bigoplus_{i=1}^n V_i$ is a Σ^{-1}-orthogonal direct sum decomposition and P_i is the Σ^{-1}-orthogonal projection onto V_i, $i = 1, \dots, n$, then the family $\big(P_i(X)\big)_{i=1}^n$ is independent.*

Proof. Given $i, j = 1, \dots, n$ with $i \neq j$ we check that

$$\mathrm{Cov}\big(P_i(X), P_j(X)\big) = 0.$$

For all $\alpha \in V_i^*$ and $\beta \in V_j^*$, we have:

$$\mathrm{Cov}\big(P_i(X), P_j(X)\big)(\alpha, \beta) = \Sigma(\alpha \circ P_i, \beta \circ P_j) = (\beta \circ P_j)\big(\Sigma(\alpha \circ P_i)\big).$$

Now note that for any $\gamma \in V^*$ the vector $\Sigma(\gamma) \in V$ is Σ^{-1}-orthogonal to the kernel of γ as $\Sigma^{-1}\big(\Sigma(\gamma), v\big) = \gamma(v) = 0$ for every $v \in \mathrm{Ker}(\gamma)$. Then $\Sigma(\alpha \circ P_i)$ is Σ^{-1}-orthogonal to the kernel of P_i and therefore $\Sigma(\alpha \circ P_i) \in V_i$. The proof is concluded by noting that P_j annihilates V_i. □

25. Expected value of quadratic forms

If X is an integrable random vector then the expected value of a linear function of X can be easily calculated by applying the same linear function to the expected value of X (see (11.1)). The next natural question is how to determine the expected value of a quadratic form in X. Recall that a *quadratic form* on a vector space V is a map of the form

$$V \ni x \longmapsto B(x, x) \in \mathbb{R},$$

where $B : V \times V \to \mathbb{R}$ is a bilinear form. It can be assumed without loss of generality that B is symmetric, but we won't do that as the formula that we will obtain holds also if B is not symmetric.

Let V be a real finite-dimensional vector space, X be a square integrable V-valued random vector, $B : V \times V \to \mathbb{R}$ be a bilinear form and let us derive a formula for the expected value of $B(X, X)$. We consider first the case in which $E(X) = 0$.

Let $(e_i)_{i=1}^n$ be a basis of V and let V^* be endowed with its dual basis $(\alpha_i)_{i=1}^n$. The matrix $(B_{ij})_{n\times n}$ that represents the linear map $B : V \to V^*$ that is identified with the bilinear form B is given by

$$B_{ij} = B(e_j, e_i), \quad i, j = 1, \ldots, n.$$

Writing $X = \sum_{i=1}^n \alpha_i(X)e_i$, we obtain

$$B(X, X) = \sum_{i=1}^n \sum_{j=1}^n B_{ji}\alpha_i(X)\alpha_j(X);$$

taking expected values on both sides of the equality and using $E(X) = 0$ we get

$$E\big(B(X, X)\big) = \sum_{i=1}^n \sum_{j=1}^n B_{ji}\Sigma(\alpha_i, \alpha_j),$$

where $\Sigma = \mathrm{Var}(X)$ denotes the variance of X. Let $(\Sigma_{ij})_{n\times n}$ denote the matrix that represents the linear map $\Sigma : V^* \to V$, so that

$$\Sigma_{ij} = \Sigma(\alpha_j, \alpha_i) = \Sigma(\alpha_i, \alpha_j), \quad i, j = 1, \ldots, n$$

and:

$$E\big(B(X, X)\big) = \sum_{j=1}^n \sum_{i=1}^n B_{ji}\Sigma_{ij} = \sum_{j=1}^n (B \circ \Sigma)_{jj} = \sum_{i=1}^n (\Sigma \circ B)_{ii}. \tag{25.1}$$

Recall that the *trace* of a linear transformation T from a finite-dimensional vector space to itself, denoted by $\mathrm{tr}(T)$, is defined as the trace (i.e., the sum of the main diagonal elements) of the matrix that represents T with respect to any basis (with the same basis being used in the domain and counter-domain of T). It is easy to prove that the trace of the representing matrix does not depend on the choice of basis.

In (25.1) we have concluded the proof of the following result.

Proposition 25.1. *Let V be a real finite-dimensional vector space and X be a V-valued square integrable random vector with variance $\Sigma \in V \otimes V$. If $E(X) = 0$ then for any bilinear form $B : V \times V \to \mathbb{R}$ we have:*

$$E\big(B(X, X)\big) = \mathrm{tr}(B \circ \Sigma) = \mathrm{tr}(\Sigma \circ B). \qquad \square$$

In order to obtain a matrix representation of $\Sigma \circ B$ to compute the trace we should multiply the matrices that represent $\Sigma : V^* \to V$ and $B : V \to V^*$ when regarded as linear transformations. Recall that the matrix representing a bilinear form is the transpose of the matrix representing the linear

transformation it is identified with. In the case of Σ the distinction between the two matrices is irrelevant as its matrix is symmetric. In the case of B, if it is not symmetric, the distinction is not irrelevant for obtaining the correct matrix representation of $\Sigma \circ B$ but it is irrelevant for the value of the trace since Σ is symmetric.

Corollary 25.2. *Let V be a real finite-dimensional vector space and X be a V-valued square integrable random vector with variance $\Sigma \in V \otimes V$. For any bilinear form $B : V \times V \to \mathbb{R}$ we have:*

$$E\big(B(X,X)\big) = \operatorname{tr}(\Sigma \circ B) + B\big(E(X), E(X)\big).$$

Proof. Apply Proposition 25.1 to $X - E(X)$ and compute the expected value of $B\big(X, E(X)\big)$ by noting that the map $B\big(\,\cdot\,, E(X)\big)$ is linear. □

The following is an interesting particular case of Corollary 25.2.

Corollary 25.3. *Let V be a real finite-dimensional vector space and X be a square integrable V-valued random vector with nondegenerate variance $\Sigma \in V \otimes V$. If $\|\cdot\|$ denotes the norm associated to the inner product Σ^{-1} then:*

$$E\big(\|X\|^2\big) = \dim(V) + \|E(X)\|^2.$$

Moreover, if $P : V \to W$ is the Σ^{-1}-orthogonal projection onto a subspace W of V then:

$$E\big(\|P(X)\|^2\big) = \dim(W) + \big\|P\big(E(X)\big)\big\|^2.$$

Proof. The first equality follows directly from Corollary 25.2 by letting B be the inner product Σ^{-1}. For the second, we let B be the bilinear form

$$B = \Sigma^{-1}(P\,\cdot\,, P\,\cdot\,)$$

which is identified with the linear transformation

$$B = P^* \circ \Sigma^{-1} \circ P, \tag{25.2}$$

where P is regarded as a map from V to V. We have that the Σ^{-1}-orthogonal projection $P : V \to V$ is self-adjoint with respect to the inner product Σ^{-1}, i.e.

$$\Sigma^{-1}(P\,\cdot\,,\,\cdot\,) = \Sigma^{-1}(\,\cdot\,, P\,\cdot\,)$$

which is equivalent to:

$$\Sigma^{-1} \circ P = P^* \circ \Sigma^{-1}. \tag{25.3}$$

From (25.2) and (25.3) we obtain

$$B = \Sigma^{-1} \circ P^2 = \Sigma^{-1} \circ P$$

and therefore Corollary 25.2 yields:

$$\begin{aligned} E\big(\|P(X)\|^2\big) = E\big(B(X,X)\big) &= \operatorname{tr}(\Sigma \circ B) + B\big(E(X), E(X)\big) \\ &= \operatorname{tr}(P) + \big\|P\big(E(X)\big)\big\|^2 = \dim(W) + \big\|P\big(E(X)\big)\big\|^2. \quad \square \end{aligned}$$

26. The basic set up of statistical modelling

Imagine we want to answer a question that can be studied through empirical research. To do this, we gather relevant data for the question at hand. These data can be obtained through a planned experiment or by collecting observations retrospectively. If we wish to study these data using statistical methods, we will regard it as being generated by some random experiment or random process. This implies that the set M of all theoretically possible values for such data should be endowed with a probability measure defined on some σ-algebra $\mathcal{B}$ of subsets of M. It is convenient to regard the data as an $(M, \mathcal{B})$-valued random object X, i.e., we assume that $X : \Omega \to M$ is a measurable map with $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space and that the probability measure defined on $\mathcal{B}$ is the distribution $\mathbb{P}_X$ of X.

Remark 26.1. The choice of probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is irrelevant and one could simply take, for instance, $(\Omega, \mathcal{A}) = (M, \mathcal{B})$ and let X be the identity map. One might feel that it would be better to simply get rid of the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ altogether, but the language of random variables and random objects is really convenient for statistics and it requires a common domain for all of them.

The probability measure $\mathbb{P}_X$ is usually unknown and the goal of statistical inference is precisely to use the observed value of the data X to learn something about $\mathbb{P}_X$. If we really knew absolutely nothing about $\mathbb{P}_X$ it wouldn't be possible to draw interesting conclusions from the data, but it is often the case that we do have some prior information on $\mathbb{P}_X$ which gives us some justification to assert that $\mathbb{P}_X$ — or some approximation of $\mathbb{P}_X$ — belongs to a certain given subset of the set $\mathrm{Prob}(M, \mathcal{B})$ of all probability measures on $(M, \mathcal{B})$. This subset is usually conveniently described in parametric form, i.e., we write it as $\left\{\mathbb{P}_X^\vartheta : \vartheta \in \Theta\right\}$, with $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ a family of probability measures on $(M, \mathcal{B})$. The goal of statistical inference can then be rephrased as learning something about the unknown *parameter* $\vartheta \in \Theta$ using the observed random data X whose probability distribution depends on ϑ in a known way. The family $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ is called a *stochastic model* for the data X and the set Θ is called the *parameter space* of the model. One typically assumes that the mapping $\Theta \ni \vartheta \mapsto \mathbb{P}_X^\vartheta \in \mathrm{Prob}(M, \mathcal{B})$ is injective, since it would obviously be impossible to use the data to distinguish between two distinct values of the parameter that correspond to the exact same probability distribution for the data.

We can assume without loss of generality that the measurable space $(\Omega, \mathcal{A})$ and the measurable map $X : \Omega \to M$ are defined in such a way that $\mathbb{P}_X^\vartheta$ is the push-forward of some probability measure $\mathbb{P}^\vartheta$ on $(\Omega, \mathcal{A})$ under the map X for every $\vartheta \in \Theta$ since, as mentioned in Remark 26.1, we can for instance simple let $(M, \mathcal{B}) = (\Omega, \mathcal{A})$ and X be the identity map.

Let us look at a few simple concrete examples of stochastic models.

Example 26.2. Suppose we want to test if a coin is biased[3], i.e., if it tends to yield more heads than tails or the other way around. To this aim, we toss it a certain number n of times and we take note of the outcomes. For the sake of shortness, let us denote the two possible outcomes by 0 and 1 instead of heads and tails. The set of all possible outcomes for this random experiment is thus the finite set $M = \{0,1\}^n$, which we endow with the σ-algebra $\mathcal{B} = \wp(M)$, as there is no reason to use a smaller σ-algebra in a finite or countable set. The $(M, \mathcal{B})$-valued random object X representing the data is thus identified with an n-tuple $(X_i)_{i=1}^n$ of random objects taking values in $\{0,1\}$ (endowed with the σ-algebra of all its subsets), with X_i being the i-th coordinate of the map X. The random object X_i gives the outcome of the i-th toss of the coin, for all $i = 1, \ldots, n$. The distribution of X_i is a probability measure on $\{0,1\}$ and it is obviously completely determined by the value of $\mathbb{P}(X_i = 1) \in [0,1]$. It seems very reasonable to assume — at least if there is no big variation in the method used for tossing the coin during the experiment — that all the random objects X_i have the same distribution. This amounts to saying that there exists $p \in [0,1]$ such that

$$\mathbb{P}(X_i = 1) = p, \tag{26.1}$$

for all $i = 1, \ldots, n$. It seems also very reasonable to assume that the family $(X_i)_{i=1}^n$ is independent, so that:

$$\mathbb{P}_X = \bigotimes_{i=1}^n \mathbb{P}_{X_i}. \tag{26.2}$$

Equalities (26.1) and (26.2) completely determine the distribution of X and thus for each $p \in [0,1]$ we have a probability measure $\mathbb{P}_X^p = \bigotimes_{i=1}^n \mathbb{P}_{X_i}^p$ on $(M, \mathcal{B})$ where $\mathbb{P}_{X_i}^p(1) = p$, for all $i = 1, \ldots, n$. We have therefore described a stochastic model $(\mathbb{P}_X^p)_{p\in[0,1]}$ for the data X with parameter space $[0,1]$. In the context of this model, the bias of the coin is related to how far away from $\frac{1}{2}$ the unknown value of the parameter p is.

Example 26.3. Suppose that we have a question with r possible (mutually exclusive) answers that we could ask to any person of a given population of N individuals (such as "in what candidate do you intend to vote for in the next election?"). Let us label the r possible answers as 1, 2, ..., r and the individuals in that population as 1, 2, ..., N. Denote by

$$A : \{1, \ldots, N\} \longrightarrow \{1, \ldots, r\}$$

the map such that $A(i)$ is the answer that the i-th individual of the population would give to the question. For each $j = 1, \ldots, r$, let N_j be the number of elements of $A^{-1}(j)$, i.e., the number of people in the population that would give the j-th answer to the question and let $p_j = \frac{N_j}{N}$ be the

[3]This is a common example in probability and statistics textbooks, though it is likely the case that biased coins do not exist in the real world. See [3] for a discussion.

proportion of people in the population that would give that answer. We would like to obtain information about the value of p_j, for all $j = 1, \ldots, r$.

Suppose that N is large, so that it is impractical and expensive to simply ask the question to all individuals of the population. Instead, we will conduct an opinion poll by selecting some random sample of n individuals from the population. The process of selecting this sample can be regarded as a random experiment and the selected sample can thus be modelled in terms of a random object $U = (U_k)_{k=1}^n$ taking values in $\{1, \ldots, N\}^n$, where U_k (the k-th coordinate of U) denotes the label of the k-th individual selected for the sample. The relevant data for obtaining information about the proportions p_j is given by the random object $X = (X_k)_{k=1}^n$ where $X_k = A(U_k)$, i.e., X_k is the answer given by the k-th individual in the sample.

We have that the distribution of X is determined by the distribution of U and the (unknown) map A, while the distribution of U depends on the method used for selecting the sample. The sample selection method which makes the statistical theory simpler — though it is rarely ever used in practice — is to select the k-th individual in a way that is independent from previous selections and such that all N individuals in the population have the same probability of being selected, for all $k = 1, \ldots, n$. Note that, in particular, the same individual might end up being selected more than once, i.e., this is an instance of *sampling with replacement*. Using this sampling method we have that the family $(U_k)_{k=1}^n$ is independent and that $\mathbb{P}(U_k = i) = \frac{1}{N}$ for all $k = 1, \ldots, n$ and all $i = 1, \ldots, N$. It then follows that $(X_k)_{k=1}^n$ is also independent and that $\mathbb{P}(X_k = j) = p_j$, for all $k = 1, \ldots, n$ and all $j = 1, \ldots, r$. The distribution $\mathbb{P}_X = \bigotimes_{k=1}^n \mathbb{P}_{X_k}$ of the data X is thus determined by the unknown parameter $p = (p_j)_{j=1}^r$, which is an r-tuple of nonnegative numbers adding to 1. The stochastic model for the data X is then a family $(\mathbb{P}_X^p)_{p \in \Theta}$ with parameter space Θ given by:

$$\Theta = \big\{p \in [0,1]^r : \textstyle\sum_{j=1}^r p_j = 1\big\}.$$

The examples above illustrate the relevance of the following definition.

Definition 26.4. Let $(M, \mathcal{B})$ be a measurable space. A family $(X_i)_{i \in I}$ of $(M, \mathcal{B})$-valued random objects on the same probability space is said to be *independent identically distributed* (abbreviated, i.i.d.) if the family $(X_i)_{i \in I}$ is independent and $\mathbb{P}_{X_i} = \mathbb{P}_{X_j}$, for all $i, j \in I$. An i.i.d. family $(X_i)_{i=1}^n$ of size n is also called a *simple random sample with replacement* (or just *simple random sample*) of size n of the probability measure on $(M, \mathcal{B})$ given by the common distribution of all X_i.

Example 26.5. Suppose that we want to test a new drug for reducing blood pressure and to this aim we are conducting a clinical trial. For simplicity of exposition we will consider a clinical trial that is not placebo controlled, i.e., we are going to enroll a certain number n of patients for the trial and all patients will receive the drug. We will compare the systolic blood pressure X_i^{b} of the i-th patient at some moment before that patient takes the drug

with the systolic blood pressure X_i^{a} of the i-th patient at some moment after that patient takes the drug, for all $i = 1, \ldots, n$. We regard X_i^{b} and X_i^{a} as random variables, so that $X_i = (X_i^{\mathrm{b}}, X_i^{\mathrm{a}})$ is an $\mathbb{R}^2$-valued random vector and $X = (X_1, \ldots, X_n)$ is a random vector taking values in $(\mathbb{R}^2)^n \cong \mathbb{R}^{2n}$.

What is a suitable model for the data X? The situation here is not as simple as it was in Examples 26.2 and 26.3. Clearly, it is not reasonable to assume that the random variables X_i^{b} and X_i^{a} are independent, as they correspond to data obtained from the same patient. On the other hand, it seems reasonable to assume that the family $(X_i)_{i=1}^n$ is independent and even that all the X_i have the same distribution, though neither assumption is as clearly justified as in the previous examples. We will discuss the reasonability of these assumptions in a moment, but for now let us just assume that the family $(X_i)_{i=1}^n$ is independent identically distributed.

What distribution should we use for X_i? One possibility is to assume nothing about such distribution. We let the parameter space Θ be the set of all probability measures on the Borel σ-algebra of $\mathbb{R}^2$ and for every $\mathcal{P} \in \Theta$ we set $\mathbb{P}_X^{\mathcal{P}} = \bigotimes_{i=1}^n \mathbb{P}_{X_i}^{\mathcal{P}}$ with $\mathbb{P}_{X_i}^{\mathcal{P}} = \mathcal{P}$ for all $i = 1, \ldots, n$. This yields a stochastic model $(\mathbb{P}_X^{\mathcal{P}})_{\mathcal{P} \in \Theta}$ for the data X with a complicated parameter space Θ. Alternatively, one could assume that a normal distribution for X_i is a reasonable approximation. This assumption leads to a model with a simpler parameter space which is the set of pairs $\vartheta = (\mu, \Sigma)$, with $\mu \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^2 \otimes \mathbb{R}^2$ symmetric and positive semi-definite. For each $\vartheta = (\mu, \Sigma)$ we would then set $\mathbb{P}_X^{\vartheta} = \bigotimes_{i=1}^n \mathbb{P}_{X_i}^{\vartheta}$, with $\mathbb{P}_{X_i}^{\vartheta}$ a normal distribution in $\mathbb{R}^2$ with mean μ and variance Σ.

Let us now do a more in depth discussion of how realistic is the assumption that $(X_i)_{i=1}^n$ is independent identically distributed. We start with the assumption that all X_i have the same distribution. Both the baseline value of systolic blood pressure X_i^{b} and the difference $X_i^{\mathrm{b}} - X_i^{\mathrm{a}}$ (which is related to the effect of the drug) are influenced by certain characteristics of the patient (such as age, genetics, other clinical conditions that the patient might have, etc). When we say that all X_i have the same distribution we do not mean that all patients have the same characteristics, of course, as this will always be false. What we do mean is that we are sampling from a hypothetic population of possible patients characteristics using the same probability distribution, for all i. More precisely, if U_i is a random object representing the relevant set of characteristics for the i-th patient, we are assuming that all the U_i have the same distribution. Thus, for example, the probability that the i-th patient is more than 40 years old should be the same as the probability that the j-th patient is more than 40 years old.

If X_i is completely determined from U_i, i.e., if there exists a fixed measurable function f such that $X_i = f(U_i)$, it will then follow that all X_i have the same distribution if all U_i have the same distribution. More realistically, we could have that X_i is not determined by U_i, but there is some residual random noise. Such random noise can be represented by a kernel K such that

if U_i attains a certain value u then we get a certain probability distribution $\mathbb{P}(X_i \in \cdot | U_i = u) = K(u)$ for X_i conditioned on $U_i = u$. If the random noise K is the same for all patients and if all U_i have the same distribution, it will again be the case that all X_i have the same distribution, as the distribution of the pair (U_i, X_i) is $\mathbb{P}_{U_i} \star K$.

Is it reasonable to assume that all the U_i have the same distribution? The answer is definitely yes if we were picking patients by random sampling from some population using a fixed probability distribution as in Example 26.3. However, that is not how clinical research is conducted, i.e., we don't go out there drawing random people from the population. In practice, patients which are actively looking for medical treatment for a certain condition will get enrolled for the study based on satisfying certain inclusion criteria and signing a consent form. It could happen, for instance, that patients with different characteristics are more likely to look for treatment during different times of the year, for example, and this would make the assumption that all U_i have the same distribution not valid.

What about independence of $(X_i)_{i=1}^n$? This seems like a reasonable assumption, but here is an example where it fails: imagine that we have many different hospitals enrolling patients and that patients from different hospitals tend to have different characteristics. Thus, if H_i is a random object corresponding to the hospital that enrolled the i-th patient, we have that X_i and H_i are not independent, i.e., the conditional distribution $\mathbb{P}(X_i \in \cdot | H_i = h)$ depends on the hospital h. Note that this is not incompatible with all X_i having the same distribution, since this will hold if all H_i have the same distribution and the conditional distributions $\mathbb{P}(X_i \in \cdot | H_i = h)$ are the same for all i. Now imagine that each hospital have limited resources, so that if one hospital enrolls too many patients at the beginning, it will enroll less patients later on. This will make the family $(H_i)_{i=1}^n$ not independent and this could lead to the family $(X_i)_{i=1}^n$ not being independent as well. For example, if only one hospital tends to enroll older patients and lots of older patients got enrolled at the beginning of the trial then probably there won't be many older patients being enrolled after that.

27. The parameters of a stochastic model

As discussed in Section 26, a stochastic model consists of a family of probability measures $(\mathbb{P}_X^\vartheta)_{\vartheta \in \Theta}$ on a measurable space $(M, \mathcal{B})$, where Θ is the so called parameter space and X is an $(M, \mathcal{B})$-valued random object representing the data. In some examples, the parameter $\vartheta \in \Theta$ is characterized by a list of real numbers of fixed finite size satisfying some condition, so that we can regard Θ as a subset of some $\mathbb{R}^n$. For instance, in Example 26.2 the parameter is just a real number between 0 and 1, so that $\Theta = [0, 1]$ and in Example 26.3 the parameter is an r-tuple of nonnegative real numbers adding to 1, so that Θ is a subset of $\mathbb{R}^r$ (which is contained in an affine subspace of dimension $r - 1$).

In Example 26.5, in the simplified model in which the data X_i for the i-th patient are assumed to be normal, the parameter is a pair $\vartheta = (\mu, \Sigma)$, with $\mu \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^2 \otimes \mathbb{R}^2$ symmetric and positive semi-definite, so the parameter space can be identified with a subset of $\mathbb{R}^5$ (as a symmetric 2×2 matrix is characterized by 3 real numbers). On the other hand, in the more complicated model in which nothing is assumed about the distribution of X_i, the parameter space is the set of all probability measures on the Borel σ-algebra of $\mathbb{R}^2$, which is a subset of the infinite-dimensional Banach space of finite countably additive signed measures on that σ-algebra. Such subset spans the entire Banach space.

A model in which the parameter space Θ can be naturally regarded as a subset of some $\mathbb{R}^n$ is usually called a *parametric model*. Thus, the models in Examples 26.2, 26.3 and the simplified model in Example 26.5 are parametric models, while the more complicated model in Example 26.5 is not parametric. This terminology might sound weird as also in the nonparametric case we're talking about a parameter, but it probably comes from old times where people wouldn't think about more abstract objects as being legitimate parameters.

Remark 27.1. Note that, since the σ-algebra $\mathcal{B}$ is usually countably generated, the set $\mathrm{Prob}(M, \mathcal{B})$ of all probability measures on $(M, \mathcal{B})$ has the cardinality of the *continuum* and therefore we can always parameterize an arbitrary subset of $\mathrm{Prob}(M, \mathcal{B})$ using a single real number. Thus, strictly speaking, any model could be regarded as a parametric model. However, in real-world applications people only care about parametric models given by mappings $\vartheta \mapsto \mathbb{P}_X^\vartheta \in \mathrm{Prob}(M, \mathcal{B})$ that are nice and have some intuitive clear meaning, which are typically mappings in which a probability density function for X can be written as a nice (say, smooth) function of ϑ. Parameterizing the set of all probability measures in the Borel σ-algebra of $\mathbb{R}^2$ using a single real number (or an n-tuple of real numbers) would involve some bizarre useless mapping which will not satisfy the assumptions of typical theorems of the theory of parametric models.

If the parameter space Θ is (identified with) a subset of $\mathbb{R}^n$, so that each $\vartheta \in \Theta$ is an n-tuple of real numbers, one will usually call the n coordinates of ϑ the various *parameters* of the model. Expressions constructed combining the coordinates of ϑ are usually also called parameters. This idea is made precise in the following definition.

Definition 27.2. If $(\mathbb{P}_X^\vartheta)_{\vartheta \in \Theta}$ is a stochastic model with parameter space Θ then a *parameter* for this model is any map θ whose domain is Θ.

A parameter θ is often real-valued, i.e., it is a map $\theta : \Theta \to \mathbb{R}$ taking values in $\mathbb{R}$. Definition 27.2 has the same spirit as the definition of random variable or random object (but without the randomness). Thus, one talks about a function $f(\theta)$ of a parameter θ meaning the composition $f \circ \theta$. The identity mapping $\theta : \Theta \to \Theta$ is the *full parameter* and every other parameter

is a function of such parameter. Let us illustrate Definition 27.2 with a few examples.

Example 27.3. In Example 26.3, the coordinates p_j of $p \in \Theta$ are parameters of the model in the sense of Definition 27.2 when we identify them with the projection maps:

$$\mathbb{R}^r \supset \Theta \ni p \longmapsto p_j \in [0,1], \quad j = 1, \dots, r.$$

The difference $p_1 - p_2$ (or any function combining the various p_j) is also an example of a parameter for that model.

Example 27.4. In the simplified model in Example 26.5, we regard μ and Σ as parameters by identifying them with the projection maps:

$$\mathbb{R}^2 \times (\mathbb{R}^2 \otimes \mathbb{R}^2) \supset \Theta \ni (\mu, \Sigma) \longmapsto \mu \in \mathbb{R}^2,$$
$$\mathbb{R}^2 \times (\mathbb{R}^2 \otimes \mathbb{R}^2) \supset \Theta \ni (\mu, \Sigma) \longmapsto \Sigma \in \mathbb{R}^2 \otimes \mathbb{R}^2.$$

In this example, the parameter which one is most likely to care about is the difference $\mu_1 - \mu_2$, where μ_1 and μ_2 are the coordinates of μ. Namely, such difference is one possible way of expressing the *effect size* of the drug being tested. In the more complicated model in Example 26.5, in which the parameter space Θ is the set of all probability measures on the Borel σ-algebra of $\mathbb{R}^2$, the corresponding parameter of interest which expresses the effect size of the drug is given by

$$\Theta \ni \mathcal{P} \longmapsto \int_{\mathbb{R}^2} (x^{\mathrm{b}} - x^{\mathrm{a}}) \, \mathrm{d}\mathcal{P}(x^{\mathrm{b}}, x^{\mathrm{a}}) \in \mathbb{R}.$$

Note that the value of this parameter coincides with the expected value (under the probability measure $\mathbb{P}^{\mathcal{P}}$) of $X_i^{\mathrm{b}} - X_i^{\mathrm{a}}$, for all $i = 1, \dots, n$.

Going back to the simplified model in Example 26.5, since $\mu_1 - \mu_2$ is the parameter of interest, one might choose to *reparameterize* the model using the bijective map

$$\Theta \ni (\mu_1, \mu_2, \Sigma) \longmapsto (\mu_1, \mu_1 - \mu_2, \Sigma) \in \Theta; \tag{27.1}$$

this means that the map $\Theta \ni \vartheta \mapsto \mathbb{P}_X^{\vartheta}$ will be replaced with its composition with the inverse of the map (27.1). This yields a new parameterization of the same set of probability measures. The advantage of this new parameterization is that the parameter of interest now appears more explicitly as one of the coordinates of the full parameter. Such coordinates can now be divided into the *parameter of interest*, which is $\mu_1 - \mu_2$, and the other parameters μ_1 and Σ which we don't really care about, but are forced to include in the model as the probability distribution of the data depends on them. Parameters which we must include in the model but are not of interest are called *nuissance parameters*.

Let us give a formal definition of reparameterization of a model.

Definition 27.5. If $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ is a stochastic model with parameter space Θ and $\varphi : \Theta' \to \Theta$ is a bijective map defined on some set Θ' then the family

$$(\mathbb{P}_X^{\varphi(\vartheta')})_{\vartheta'\in\Theta'}$$

obtained by taking the composition of $\Theta \ni \vartheta \mapsto \mathbb{P}_X^\vartheta$ with φ is called a *reparameterization* of the model $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$.

For practical purposes, a reparameterization of a model is really the same model in the sense that both ways of modelling the data involve the same assumptions and approximations; we are only relabeling the parameters in a possibly more convenient way.

28. The fundamental ideas of statistical inference

Consider a stochastic model $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ for some data X. We use the following mathematical set up: $(\Omega, \mathcal{A})$ and $(M, \mathcal{B})$ are measurable spaces, $X : \Omega \to M$ is a measurable map, $(\mathbb{P}^\vartheta)_{\vartheta\in\Theta}$ is a family of probability measures on $(\Omega, \mathcal{A})$ and $\mathbb{P}_X^\vartheta$ is the push-forward of $\mathbb{P}^\vartheta$ under X, i.e., $\mathbb{P}_X^\vartheta$ is the distribution of the random object X when its domain is endowed with the probability measure $\mathbb{P}^\vartheta$. As mentioned in Remark 26.1, the choice of $(\Omega, \mathcal{A})$ is not important.

As discussed in Section 26, the goal of statistical inference is to obtain information about the true value of the unknown parameter $\vartheta \in \Theta$ using the observed known value of the data X. In many situations, we are not interested in ϑ itself, but only on the value of $\theta(\vartheta)$ for some map θ defined on Θ. Recall that such a map θ is called *a parameter* of the model (Definition 27.2). Clearly, there cannot be a procedure that allows one to determine the exact true value of θ with certainty using the data X and the information about θ that one is going to obtain will involve error margins and probabilities.

The type of question that most normal people — that is, people who don't have formal training in statistics — tend to ask in the context of problems involving statistical inference are questions regarding probabilities for the value of a parameter θ. For instance, consider Example 26.3 and assume that we are dealing with electoral polling. Most people would typically ask a question like "what is the probability that this candidate would win the election if it happened today?". That is equivalent to asking for the probability that $p_1 > p_j$ for all $j = 2, \ldots, r$, where 1 is the number associated to the candidate in which one is interested. Regarding Example 26.5, most people would like to know "the probability that the drug works" or maybe the probability that the effect size of the drug is larger than some number. In both cases, the questions being asked involve probabilities for the values of parameters.

Unfortunately, within the mathematical set up that we are considering, probabilities involving parameters are simply meaningless, as parameters are not random objects. The parameter space Θ in which parameters are

defined is simply a set, it is not endowed with a probability measure. Within the domain of so called *classical* or *frequentist* statistics, probabilities about parameters are indeed regarded as meaningless and only probabilities about outcomes of random experiments are meaningful. Recall that such probabilities are interpreted as limits of frequencies (see the discussion in Section 2). Answers to questions regarding probabilities about parameters are the subject of *Bayesian* statistics which we will discuss at the end of the section.

Frequentist statistics presents us with a peculiar situation: once the data X is observed, we are interested is saying something about the value of some unknown parameter θ. As we cannot determine the value of θ with certainty, the only type of statement that we could possibly be able to make about θ are probabilistic statements and yet probabilities about the value of θ are meaningless. On the other hand, probabilities about X are meaningful but, since X is known, what is the purpose of making probabilistic statements about X?

Let us explain how this conundrum is resolved. Briefly speaking, classical frequentist statistics tell us how often we will make a mistake if we behave in a certain way after we look at the data. More specifically, it says how often we will make a mistake (i.e., what is the probability that we will make a mistake) if we use a certain procedure to make statements about θ based on the observed value of X. Since the statement we will make about θ is a function of X, it inherits the randomness from X and it becomes a *random statement*, so that it makes sense to ask for the probability that such statement about θ will be wrong even though θ is not itself random.

This is the picture that a practitioner of classical frequentist statistics should have in mind: the situation you are going through right now (observing X and saying something about θ) is one of a potential large number of similar instances in which some data X is generated. The frequencies for the values of X are given by the probability measure $\mathbb{P}_X^{\vartheta}$ for some unknown $\vartheta \in \Theta$ that is fixed throughout the various repetitions of the process that generates X. If we make a statement about $\theta(\vartheta)$ using some procedure that is based on the data X, then such statement will vary across iterations of the process because the data X varies. As our statement varies and the value of θ is kept fixed, there will be lucky iterations in which our statement is correct and unlucky iterations in which our statement is wrong. What we want is to calculate the probability (i.e., the frequency) with which we will make wrong statements and we want to use this ability to calculate such probabilities to calibrate the statement-generating procedures in such a way that we will rarely make wrong statements. The notion of "rarely" is of course vague: depending on the issue at hand, you might be happy to be wrong at most 10% of the time, while for other issues you might think that being wrong 0.5% of the time is the maximum error rate that is tolerable.

What kinds of statements can be made about the value of θ? Since probabilistic statements are meaningless, the only type of statement that can be made is of the form "the value of θ belongs to C", where C is some

subset of the counterdomain of the map θ. Such set C is obtained from the data X (i.e., it is a function of X) and thus it is a *random set*. One can then try to design a set C that depends on X in such a way that, no matter what the true value of θ is, there is some upper bound on how often C will not contain the value of θ (i.e., how often the statement $\theta(\vartheta) \in C$ will be wrong). This leads to the notion of *confidence set* which we discuss in more detail in Section 29. Such sets are typically intervals and that is why they are more usually known as *confidence intervals*. One might also be interested in upper bounds on error rates that are dependent on the value of θ. For instance, you might have a certain upper bound for the error rate when the value of θ satisfies a certain hypothesis and another upper bound for the error rate when the value of θ satisfies a different hypothesis. This leads to the idea of *hypothesis testing* which we discuss in detail in Section 31.

28.1. **Bayesian statistics.** As discussed above, Bayesian inference yields answers to questions regarding probabilities for the parameters of a model. Thus, in order to do Bayesian inference we need to adapt our mathematical formalism in such a way that parameters become random objects, i.e., we need to add to our formalism a probability measure defined on some σ-algebra of subsets of the parameter space Θ. How should such probability measure be interpreted?

In Bayesian statistics, probabilities are not necessarily related to frequencies of occurrences of events when some experiment is repeated. Most often, probabilities are interpreted as a quantitative expression of (usually imperfect) knowledge regarding the state of the world. The Dirac delta probability measure is the extreme case which corresponds to perfect knowledge about a certain quantity and probability measures that are highly concentrated in small sets express a high amount of knowledge about the value of that quantity. On the other hand, probability measures that are very spread out express the idea of very little knowledge about the value of the quantity.

The fundamental structure of Bayesian inference is the following: one starts with a *prior distribution* on the parameter space Θ, which is a probability measure in which we try to encode what is known about the parameters before the recently gathered data X is observed. We then use the data X to update the prior distribution obtaining a *posterior distribution* on the parameter space. In mathematical terms, we endow the parameter space Θ with a σ-algebra $\mathcal{A}_\Theta$ and a probability measure $\mathbb{P}_\Theta$ defined on that σ-algebra. Such probability measure is the prior distribution. The σ-algebra $\mathcal{A}_\Theta$ should be such that the stochastic model $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ becomes a kernel

$$\Theta \ni \vartheta \longmapsto \mathbb{P}_X^\vartheta \in \mathrm{Prob}(M, \mathcal{B})$$

with source $(\Theta, \mathcal{A}_\Theta)$ and target $(M, \mathcal{B})$. In other words, for every $B \in \mathcal{B}$, the map $\vartheta \mapsto \mathbb{P}_X^\vartheta(B)$ should be measurable with respect to $\mathcal{A}_\Theta$. Without loss of generality, we can assume that

$$\Theta \ni \vartheta \longmapsto \mathbb{P}^\vartheta \in \mathrm{Prob}(\Omega, \mathcal{A}) \tag{28.1}$$

is a kernel with source $(\Theta, \mathcal{A}_\Theta)$ and target $(\Omega, \mathcal{A})$. Taking the star product of the prior distribution $\mathbb{P}_\Theta$ with the kernel (28.1) we obtain a probability measure on the σ-algebra $\mathcal{A}_\Theta \otimes \mathcal{A}$ of subsets of the cartesian product $\Theta \times \Omega$. We can now regard all random objects defined on $(\Omega, \mathcal{A})$ as random objects defined on $(\Theta \times \Omega, \mathcal{A}_\Theta \otimes \mathcal{A})$ by identifying them with their compositions with the second projection of $\Theta \times \Omega$. Moreover, whenever $\theta : \Theta \to \Theta'$ is a parameter and the set Θ' is endowed with a σ-algebra for which the map θ is measurable, we can regard θ as a random object defined on $(\Theta \times \Omega, \mathcal{A}_\Theta \otimes \mathcal{A})$ by taking its composition with the first projection.

Once the constructions above are performed, everything that used to be a random object in the frequentist (prior free) setting remains a random object. In addition, every (measurable) parameter is now also a random object and all these random objects are defined on the same probability space. The *posterior distribution* for a parameter θ is now obtained by conditioning θ on the observed data X. More precisely, we consider a regular conditional distribution

$$\mathbb{P}(\theta \in \cdot \,|X = x) \tag{28.2}$$

of θ given X and we replace $x \in M$ with the data that was observed. The posterior distribution for θ should be interpreted as an update of our knowledge about the value of θ after the data was observed. Such posterior distribution can now be used to answer questions about the probability of the value of θ belonging to a certain (measurable) set. Thus, one can give an answer for the probability that a certain candidate would win the election today after an electoral pool is conducted and one can tell what is the probability that a certain drug works after performing a clinical trial.

There is a small technical problem that we have to handle. Recall that the regular conditional distribution of θ given X is only defined up to $\mathbb{P}_X$-almost everywhere equality and thus, unless the observed data $x \in M$ is such that $\mathbb{P}(X = x) > 0$, the value of (28.2) is meaningless. This problem is solved by adding some natural requirement for the regular conditional probability in order for it to have a canonical representative. In Subsection 28.2 we have discussed the uniqueness problem for regular conditional distributions and we have shown that, under mild conditions, the value of (28.2) is uniquely defined if x belongs to the support of $\mathbb{P}_X$ and the regular conditional distribution of θ given X is required to be continuous with respect to the weak topology at the point x. In concrete problems, a continuous regular conditional probability usually exists and thus there is no real difficulty here.

Critics of Bayesian methods usually complain about the fact that the choice of prior distribution on the parameter space is arbitrary and cannot be justified. Such criticism does have some merit and choosing an appropriate prior and giving a justification for it might be a hard problem. Though we often do have some prior information on parameters, it is hard to translate such information into a concrete probability measure. So although Bayesian statistics gives answers that are more natural and understandable

to normal people, the difficulty with choosing the prior is a price to pay. As a reply to critics of Bayesian methods, it could be pointed out that often frequentist methods do suffer from similar difficulties as usually there are lots of somewhat arbitrary choices that are hard to justify involving model and inference method selection.

29. Confidence sets

The motivation for confidence sets was discussed in Section 28 and now we proceed to the technical details. We start by defining the notion of a random set. Generally speaking, in probability theory, a random "something" is a measurable map defined on a probability space taking values in the set where "something" belongs. In particular, a random set is a measurable map defined on a probability space taking values in a set of sets. In order for the notion of measurability to make sense in this context, we need to endow the set $\wp(\Theta)$ of all subsets of a set Θ with a σ-algebra.

Definition 29.1. Given a set Θ, the *canonical σ-algebra* of subsets of $\wp(\Theta)$ is the σ-algebra induced by the family of maps

$$\delta_\theta : \wp(\Theta) \ni A \longmapsto \mathbf{1}_A(\theta) \in \{0,1\}, \quad \theta \in \Theta,$$

where $\{0,1\}$ is endowed with the σ-algebra $\wp\big(\{0,1\}\big)$.

Recall that the set $\wp(\Theta)$ of all subsets of Θ is naturally identified with the cartesian product $\{0,1\}^\Theta = \prod_{\theta\in\Theta}\{0,1\}$ by associating each subset A of Θ with its indicator function $\mathbf{1}_A : \Theta \to \{0,1\}$. Under such identification, the map δ_θ is identified with the projection onto the θ-th coordinate and thus the canonical σ-algebra of subsets of $\wp(\Theta)$ is identified with the product σ-algebra on $\{0,1\}^\Theta$, where each factor $\{0,1\}$ is endowed with the σ-algebra of all its subsets.

From now on, the set $\wp(\Theta)$ will always be assumed to be endowed with its canonical σ-algebra, unless explicitly stated otherwise.

Definition 29.2. Given a set Θ, by a *random subset of* Θ we mean a measurable map C defined on some probability space taking values in $\wp(\Theta)$. By a *random set* we mean a random subset of Θ for some set Θ.

The following is a straightforward consequence of the fundamental property of the σ-algebra induced by a family of maps.

Proposition 29.3. *Let $(M,\mathcal{B})$ be a measurable space and Θ be a set. We have that a map $C : M \to \wp(\Theta)$ is measurable if and only if the set*

$$\big\{x \in M : \theta \in C(x)\big\} \tag{29.1}$$

is in $\mathcal{B}$, for all $\theta \in \Theta$. □

Using the measurability criteria given in Proposition 29.3 one immediately sees that replacing Θ with a different set that contains $C(x)$ for all $x \in M$ does not affect the measurability of C.

Corollary 29.4. *Let $(M, \mathcal{B})$ be a measurable space, Θ be a set and Θ' be a set containing Θ. We have that a map $C : M \to \wp(\Theta)$ is measurable if and only if it is measurable when regarded as a $\wp(\Theta')$-valued map.* □

In previous sections we used to write $\mathbb{P}(X \in B)$ for the probability that (the value of) a random object X belongs to a fixed (nonrandom) set B. Now that we have defined random sets we can turn things around and consider the probability that a fixed (nonrandom) object belongs to a random set.

Definition 29.5. Let $(M, \mathcal{B}, \mathbb{P})$ be a probability space, Θ be a set and let $C : M \to \wp(\Theta)$ be a random set. We write

$$\mathbb{P}(\theta \in C) = \mathbb{P}\big(\big\{x \in M : \theta \in C(x)\big\}\big),$$

for every $\theta \in \Theta$.

The fact that the probability $\mathbb{P}(\theta \in C)$ is well-defined follows from the measurability of the set (29.1).

A convenient way to visualize a map $C : M \to \wp(\Theta)$ is to identify it with the subset $\bigcup_{x \in M} \big(C(x) \times \{x\}\big)$ of the cartesian product $\Theta \times M$ and to imagine such product as a rectangle with Θ in the horizontal axis and M in the vertical axis. Having this in mind, we introduce the following terminology.

Definition 29.6. Let M and Θ be sets and let C be a subset of $\Theta \times M$. For every $\theta \in \Theta$ we define the *θ-th column* of C by

$$C_\theta = \big\{x \in M : (\theta, x) \in C\big\}$$

and for every $x \in M$ we define the *x-th row* of C by:

$$C^x = \big\{\theta \in \Theta : (\theta, x) \in C\big\}.$$

The map

$$M \ni x \longmapsto C^x \in \wp(\Theta)$$

is called the *row map* of C and

$$\Theta \ni \theta \longmapsto C_\theta \in \wp(M)$$

is called the *column map* of C.

Clearly:

$$x \in C_\theta \iff (\theta, x) \in C \iff \theta \in C^x,$$

for all $\theta \in \Theta$ and all $x \in M$. The equalities

$$C = \bigcup_{\theta \in \Theta} \big(\{\theta\} \times C_\theta\big) = \bigcup_{x \in M} \big(C^x \times \{x\}\big)$$

show that a subset C of $\Theta \times M$ is uniquely determined by either its row map or by its column map.

The following is a restatement of Proposition 29.3 in terms of subsets of $\Theta \times M$.

Proposition 29.7. *Let $(M, \mathcal{B})$ be a measurable space, Θ be a set and let C be a subset of $\Theta \times M$. We have that the row map of C is measurable if and only if every column of C is measurable (i.e., $C_\theta \in \mathcal{B}$ for all $\theta \in \Theta$).* □

We now have all the requisites for the definition of a confidence set. We consider the same mathematical set up as in Section 28.

Definition 29.8. Let $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ be a stochastic model, with the data X taking values in a measurable space $(M, \mathcal{B})$, and let $\theta : \Theta \to \Theta'$ be a parameter for the model. Given $\gamma \in [0, 1]$, by a *γ-confidence set for the parameter θ* we mean a measurable map $C : M \to \wp(\Theta')$ such that the random set $C(X)$ satisfies

$$\mathbb{P}^\vartheta\big(\theta(\vartheta) \in C(X)\big) \geq \gamma,$$

for all $\vartheta \in \Theta$. We say that C is a *γ-confidence set for the parameter θ in the strict sense* if the equality

$$\mathbb{P}^\vartheta\big(\theta(\vartheta) \in C(X)\big) = \gamma$$

holds for all $\vartheta \in \Theta$.

The number γ is usually called the *confidence level*. The value $\gamma = 0.95$ is a very popular choice of confidence level in scientific papers, but of course this is an arbitrary convention.

Recall that the probability measure $\mathbb{P}_X^\vartheta$ is the push-forward of $\mathbb{P}^\vartheta$ under the map X and therefore:

$$\mathbb{P}^\vartheta\big(\theta(\vartheta) \in C(X)\big) = \mathbb{P}_X^\vartheta\big(\theta(\vartheta) \in C\big). \tag{29.2}$$

The following result is an immediate consequence of (29.2), Proposition 29.7 and the definition of confidence set.

Proposition 29.9. *Let $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ be a stochastic model, with the data X taking values in a measurable space $(M, \mathcal{B})$, and let $\theta : \Theta \to \Theta'$ be a parameter for the model. Given $\gamma \in [0, 1]$, we have that a γ-confidence set for the parameter θ (resp., a γ-confidence set for the parameter θ in the strict sense) is the same thing as the row map of a subset C of $\Theta' \times M$ such that every column of C is measurable and such that, for all $\vartheta \in \Theta$, the probability of the $\theta(\vartheta)$-th column of C with respect to the probability measure $\mathbb{P}_X^\vartheta$ is greater than or equal to γ (resp., equal to γ).* □

Proposition 29.9 gives us the following recipe for constructing all possible γ-confidence sets (resp., all possible γ-condifence sets in the strict sense) for a parameter $\theta : \Theta \to \Theta'$: for each $\vartheta' \in \Theta'$ in the image of θ, choose a measurable subset $C_{\vartheta'}$ of M such that the probability $\mathbb{P}_X^\vartheta(C_{\vartheta'})$ is greater than or equal to γ (resp., equal to γ) for every $\vartheta \in \theta^{-1}(\vartheta')$. If there are any $\vartheta' \in \Theta'$ outside the image of θ, simply let $C_{\vartheta'}$ be an arbitrary measurable subset of M. The desired γ-confidence set is then the row map of the subset C of $\Theta' \times M$ whose ϑ'-th column is the chosen set $C_{\vartheta'}$, for all $\vartheta' \in \Theta'$. In particular, if $\theta : \Theta \to \Theta$ is the identity map (what we call the full parameter of the model) then specifying a γ-confidence set (resp., a γ-confidence set in

the strict sense) for θ is equivalent to choosing for every $\vartheta \in \Theta$ a measurable subset C_ϑ of M such that $\mathbb{P}_X^\vartheta(C_\vartheta)$ is greater than or equal to γ (resp., equal to γ).

Do γ-confidence sets always exist? Trivially, if we set $C(x) = \Theta'$ for all $x \in M$ then $C : M \to \wp(\Theta')$ will be a γ-confidence set for every $\gamma \in [0,1]$ and every parameter $\theta : \Theta \to \Theta'$. However, for a given confidence level γ, one would normally prefer to define a confidence set that is, in some sense, as small as possible and therefore a γ-confidence set in the strict sense is usually preferred over a γ-confidence set C for which the probability $\mathbb{P}_X^\vartheta\big(\theta(\vartheta) \in C\big)$ can be greater than γ.

There are two obstructions for the existence of γ-confidence sets in the strict sense. First, it may happen that for some $\vartheta \in \Theta$ there are no subsets of M whose probability with respect to $\mathbb{P}_X^\vartheta$ is equal to γ. For example, if M is finite then the image of any probability measure on $(M, \mathcal{B})$ is also finite and therefore subsets of probability γ will not exist for most γ. Even when there is no difficulty with the existence of subsets of probability γ, there is another obstruction if the parameter θ is not injective. Namely, obtaining a γ-confidence set for θ in the strict sense requires that we find a set whose probability is equal to γ with respect to multiple distinct probability measures. More specifically, for every $\vartheta' \in \Theta'$ in the image of θ, we need a measurable subset $C_{\vartheta'}$ of M such that $\mathbb{P}_X^\vartheta(C_{\vartheta'})$ is equal to γ for every ϑ in $\theta^{-1}(\vartheta')$ and such subset might not exist even if for each individual probability measure $\mathbb{P}_X^\vartheta$ there exists a subset with probability γ.

In addition to purely mathematical issues, there are other challenges related to confidence sets. As illustrated in Example 29.11 below, some confidence sets are not useful for practical purposes. Furthermore, when it comes to real-world applications, there are also computational hurdles to overcome. Simply providing an abstract definition of a confidence set is not sufficient; we must be able to write code that enables a computer to efficiently compute a reliable approximation of the set within a reasonable timeframe.

Example 29.10. Consider the stochastic model which states that X is a normal random variable with some unknown mean $\mu \in \mathbb{R}$ and some fixed known variance $\sigma^2 > 0$. More precisely, we let $(M, \mathcal{B})$ be the real line endowed with its Borel σ-algebra and $(\mathbb{P}_X^\mu)_{\mu \in \mathbb{R}}$ be the family such that $\mathbb{P}_X^\mu$ is a normal distribution with mean μ and variance σ^2. Note that if Z is a standard normal random variable defined on some probability space then

$$\mathbb{P}^\mu\Big(\frac{X-\mu}{\sigma} \in B\Big) = \mathbb{P}(Z \in B),$$

for every Borel subset B of $\mathbb{R}$. For each $\alpha \in \,]0,1[$, denote by $z_\alpha \in \mathbb{R}$ the unique real number such that $\mathbb{P}(Z > z_\alpha) = \mathbb{P}(Z < -z_\alpha) = \alpha$. We have

$$\mathbb{P}(-z_{\frac{1-\gamma}{2}} \leq Z \leq z_{\frac{1-\gamma}{2}}) = \gamma,$$

for every $\gamma \in]0,1[$ and thus

$$\mathbb{P}^{\mu}(\mu - \sigma z_{\frac{1-\gamma}{2}} \leq X \leq \mu + \sigma z_{\frac{1-\gamma}{2}}) = \gamma,$$

for every $\mu \in \mathbb{R}$. In other words, the interval

$$[\mu - \sigma z_{\frac{1-\gamma}{2}}, \mu + \sigma z_{\frac{1-\gamma}{2}}] \tag{29.3}$$

has probability γ with respect to the probability measure $\mathbb{P}^{\mu}_{X}$ for all $\mu \in \mathbb{R}$. Let C be the subset of $\mathbb{R}^2$ whose μ-th column is (29.3) for every $\mu \in \mathbb{R}$. The row map of C is given by

$$\mathbb{R} \ni x \longmapsto C^x = [x - \sigma z_{\frac{1-\gamma}{2}}, x + \sigma z_{\frac{1-\gamma}{2}}] \in \wp(\mathbb{R}) \tag{29.4}$$

and therefore (29.4) is a γ-confidence set for the parameter μ in the strict sense.

Example 29.11. Let $(\mathbb{P}^{\vartheta}_{X})_{\vartheta \in \Theta}$ be a stochastic model, with the data X taking values in a measurable space $(M, \mathcal{B})$, and let $\theta : \Theta \to \Theta'$ be a parameter for the model. If $C : M \to \wp(\Theta')$ is a γ-confidence set for θ in the strict sense for a certain $\gamma \in [0,1]$ then

$$M \ni x \longmapsto \Theta' \setminus C(x) \in \wp(\Theta')$$

is a $(1-\gamma)$-confidence set for θ in the strict sense. In particular, considering the model in Example 29.10, we have that

$$\mathbb{R} \ni x \longmapsto \mathbb{R} \setminus [x - \sigma z_{\frac{\gamma}{2}}, x + \sigma z_{\frac{\gamma}{2}}] \in \wp(\mathbb{R}) \tag{29.5}$$

is a γ-confidence set in the strict sense for μ for every $\gamma \in]0,1[$. This is a correct, yet horrible confidence set for μ for most practical applications. For example, assume that the known standard deviation $\sigma > 0$ is very small. In this case, the value of X is likely to fall near μ and thus if we observe $X = x$ we should regard x as an estimate of μ and have some level of confidence that the unknown value of μ belongs to a small interval centered at x. That is the kind of confidence set we want for μ and that is precisely what (29.4) is. The set (29.5), on the other hand, is the complement of a tiny neighborhood of x. While it is true that we will be correct with frequency γ if we assert that μ is outside such tiny neighborhood of x whenever we observe $X = x$, this is typically not very useful information about μ. For example, a research paper could say that the average height of the adult female in the United States is between 1.62m and 1.64m (with 95% confidence) and that is an interesting useful conclusion, but no one would care about a paper which states that the average height of the adult female in the United States is *not* between 1.62968m and 1.63032m (with 95% confidence). This serves to illustrate the fact that merely satisfying the mathematical requirements for a confidence set is usually not all that we want.

Example 29.12. The assumption in Example 29.10 that the variance of X be known is good for illustrative simple examples in statistics textbooks, but never valid in practice. So consider the stochastic model $(\mathbb{P}^{\vartheta}_{X})_{\vartheta \in \Theta}$ in

which $\Theta = \mathbb{R} \times]0, +\infty[$ and for every $\vartheta = (\mu, \sigma) \in \Theta$ we have that $\mathbb{P}_X^\vartheta$ is a normal distribution on $\mathbb{R}$ with mean μ and variance σ^2. Specifying a γ-confidence set for μ in the strict sense is equivalent to choosing, for each $\mu \in \mathbb{R}$, a Borel subset C_μ of $\mathbb{R}$ such that $\mathbb{P}^{(\mu,\sigma)}(X \in C_\mu) = \gamma$ for all $\sigma > 0$. Except for trivial uninteresting cases[4], the probability $\mathbb{P}^{(\mu,\sigma)}(X \in C_\mu)$ is highly dependent on σ and thus there is no γ-confidence set for μ in the strict sense. This shouldn't be surprising: if we sample a single element X from $N(\mu, \sigma^2)$ then there is just no data that could possibly be used to estimate the unknown variance σ^2 and without some information about the variance we can't estimate how far from μ the value of X might be. Note that a γ-confidence set in the strict sense for the full parameter (μ, σ) can be obtained. Namely, if $\gamma \in]0, 1[$ then for each $(\mu, \sigma) \in \Theta$ the interval (29.3) has probability γ with respect to the probability measure $\mathbb{P}_X^{(\mu,\sigma)}$ and thus the desired confidence set is obtained as the row map of the subset C of $\Theta \times \mathbb{R}$ whose (μ, σ)-th column is (29.3), for all $(\mu, \sigma) \in \Theta$. Such confidence set is given by:

$$\mathbb{R} \ni x \longmapsto C^x = \left\{(\mu, \sigma) \in \Theta : |\mu - x| \leq \sigma z_{\frac{1-\gamma}{2}}\right\} \in \wp(\Theta).$$

We have that C^x is an unbounded triangular region on the half-plane Θ which is symmetrical around the axis $\{x\} \times]0, +\infty[$. Moreover, the confidence level γ determines the slope of the sides of the region.

29.1. **Functions of parameters.** Let $(\mathbb{P}_X^\vartheta)_{\vartheta \in \Theta}$ be a stochastic model, with the data X taking values in a measurable space $(M, \mathcal{B})$, and let $\theta : \Theta \to \Theta'$ be a parameter for the model. Let $f : \Theta' \to \Lambda$ be a map taking values in some set Λ and assume that we have a γ-confidence set $C : M \to \wp(\Theta')$ for the parameter θ. An obvious strategy for defining a γ-confidence set for the parameter $f(\theta) = f \circ \theta$ is to consider the map $f[C] : M \to \wp(\Lambda)$ defined by:

$$M \ni x \longmapsto f[C(x)] \in \wp(\Lambda).$$

Surely if we are γ-confident that the value of θ belongs to C then we should be γ-confident that the value of $f(\theta)$ belongs to $f[C]$, right? The short answer is "yes", but there are some difficulties, the first being very technical and the second related to practical applications.

The technical difficulty is related to measurability. The statement that $f[C]$ be a γ-confidence set for $f(\theta)$ is equivalent to the requirement that for all $\lambda \in \Lambda$, the set

$$\left\{x \in M : \lambda \in f[C(x)]\right\} \tag{29.6}$$

[4]One such trivial uninteresting case is $C_\mu = [\mu, +\infty[$, which satisfies the equality $\mathbb{P}^{(\mu,\sigma)}(X \in C_\mu) = \gamma$ for all $\sigma > 0$ if $\gamma = \frac{1}{2}$. This yields that $C^x =]-\infty, x]$ is a $\frac{1}{2}$-confidence set for μ in the strict sense. It can actually be proven that a γ-confidence set for μ in the strict sense exists if and only if $\gamma \in \left\{0, \frac{1}{2}, 1\right\}$.

be measurable and that its probability with respect to $\mathbb{P}_X^\vartheta$ be greater than or equal to γ, for every $\vartheta \in \theta^{-1}\big(f^{-1}(\lambda)\big)$. The set (29.6) can be written as

$$\big\{x \in M : \lambda \in f[C(x)]\big\} = \bigcup_{\vartheta' \in f^{-1}(\lambda)} \big\{x \in M : \vartheta' \in C(x)\big\}, \tag{29.7}$$

for all $\lambda \in \Lambda$. The fact that C is a γ-confidence set for θ says that

$$\big\{x \in M : \vartheta' \in C(x)\big\} \tag{29.8}$$

is measurable for all $\vartheta' \in \Theta'$ and that the probability of (29.8) with respect to $\mathbb{P}_X^\vartheta$ is greater than or equal to γ, for all $\vartheta \in \theta^{-1}(\vartheta')$. Since the union in (29.7) might be uncountable, there is no guarantee that the set (29.6) will be measurable, even though each term in the union is measurable.

Although artificial examples in which (29.6) fails to be measurable can be easily constructed, they do not seem to occur in practice. Moreover, under mild assumptions it can be shown that (29.6) is close to being measurable in the sense that it becomes measurable when we consider the completion of the probability measures (see Remark 29.13 below).

As it is to be expected, if (29.6) is measurable for all $\lambda \in \Lambda$ then $f[C]$ is indeed a γ-confidence set for $f(\theta)$. Namely, for all $\vartheta \in \theta^{-1}\big(f^{-1}(\lambda)\big)$ the set (29.6) contains (29.8) for $\vartheta' = \theta(\vartheta)$ and the probability of (29.8) with respect to $\mathbb{P}_X^\vartheta$ is greater than or equal to γ.

The second difficulty that arises when one uses the confidence set $f[C]$ for $f(\theta)$ is that in some cases it might be too large, so large that it is completely useless. For instance, in Example 29.12 we obtained a confidence set C for the full parameter (μ, σ) of the model. The parameters μ and σ are then functions of (μ, σ), namely, μ and σ are obtained from (μ, σ) by applying the projections. However, for all $x \in \mathbb{R}$, the first projection of C^x is equal to the entire real line $\mathbb{R}$ and the second projection of C^x is equal to the entire half line $]0, +\infty[$, so that such confidence sets yield no information at all.

Remark 29.13. A subset of a standard Borel space $(M, \mathcal{B})$ is called *analytic* if it is the image of an $(M, \mathcal{B})$-valued measurable map defined in some standard Borel space. While it is not true in general that an analytic set is measurable (meaning that it might not belong to $\mathcal{B}$), it follows from Choquet Capacitability Theorem (see [6, Theorem 4.10.12]) that every analytic set is *universally measurable* in the sense that it is measurable with respect to the completion of any finite countably additive measure defined on $\mathcal{B}$ (see also [2, Theorem 2.2.12] for a proof that does not mention capacities). Now consider a stochastic model $(\mathbb{P}_X^\vartheta)_{\vartheta \in \Theta}$ in which X takes values in a standard Borel space $(M, \mathcal{B})$, a parameter $\theta : \Theta \to \Theta'$ for the model and a map $f : \Theta' \to \Lambda$. Assume that:

- Θ' is endowed with a σ-algebra $\mathcal{A}_{\Theta'}$ and Λ is endowed with a σ-algebra $\mathcal{A}_\Lambda$ such that $(\Theta', \mathcal{A}_{\Theta'})$ and $(\Lambda, \mathcal{A}_\Lambda)$ are standard Borel spaces and f is measurable;

- C is a confidence set for θ which is the row map of a subset of $\Theta' \times M$ (also denoted by C) that belongs to the product σ-algebra $\mathcal{A}_{\Theta'} \otimes \mathcal{B}$.

Under these assumptions, we have that the set (29.6) is analytic and thus universally measurable for every $\lambda \in \Lambda$. Namely, the set (29.6) is the λ-th column of the image of $C \subset \Theta' \times M$ under the map:

$$f \times \mathrm{Id} : \Theta' \times M \ni (\vartheta', x) \longmapsto \big(f(\vartheta'), x\big) \in \Lambda \times M.$$

Since $f \times \mathrm{Id}$ is measurable, we have that $(f \times \mathrm{Id})[C]$ is analytic. As the inverse image of an analytic set by a measurable map is analytic, it follows that the columns of an analytic set are analytic and the conclusion is obtained.

30. Estimators

We consider again the mathematical set up of Section 28. Given a parameter $\theta : \Theta \to \Theta'$ for the model $(\mathbb{P}_X^\vartheta)_{\vartheta \in \Theta}$ one would typically like to find an estimate for the unknown value of θ using the observed value of the data X. So, for example, after conducting an opinion poll we would like to have an estimate of the proportion of people in the population that would give a certain answer to the question asked in the poll and after conducting a clinical trial we would like to have an estimate of the effect size of the drug.

An estimate for the value of θ should be a specific point of Θ', computed using the data X, that is in some sense likely to be close to the true unknown value of θ. An estimate for the value of θ of this type is called a *point estimate*. A point estimate should be distinguished from a *set estimate*, which is just another name for what we call a confidence set.

Usually a research paper using statistical methods to answer a question will present in its conclusion both a point estimate for some parameter and a confidence set for that parameter, which is typically an interval around the point estimate (or some other type of connected neighborhood of the point estimate in case the parameter space is multidimensional). As mentioned in Section 28, when the confidence set is an interval it is called a *confidence interval*. A confidence interval is also popularly refereed to as an *error margin* for the point estimate.

As alluded to above, a point estimate for the value of θ should be some function of the observed data X. This makes the following standard definition relevant now.

Definition 30.1. A *statistic* for a stochastic model $(\mathbb{P}_X^\vartheta)_{\vartheta \in \Theta}$ is any measurable function of the data X.

An *estimator* for a parameter $\theta : \Theta \to \Theta'$ of a stochastic model $(\mathbb{P}_X^\vartheta)_{\vartheta \in \Theta}$ is a statistic for the model taking values in Θ'. The set Θ' should be endowed with a σ-algebra, so that the measurability requirement of a Θ'-valued function makes sense. There is no precise mathematical definition for the notion of an estimator. When we say that a certain Θ'-valued statistic is an estimator for θ we are simply expressing an intention to use it for estimating the value of θ. An estimator for θ is typically denoted by $\hat{\theta}$.

Though the notion of estimator itself is not precisely defined, there are various desirable conditions that an estimator $\hat{\theta}$ for θ should satisfy that are amenable to precise mathematical formulation. One such condition is that the random object $\hat{\theta}$ should be, in a sense to be specified, close to the constant $\theta(\vartheta)$, where ϑ is the true value of the full parameter of the model.

In what follows, we assume for simplicity that the parameter θ is real-valued (i.e., $\Theta' = \mathbb{R}$, endowed with its Borel σ-algebra), though it would be easy to generalize the following considerations to a parameter taking values in a real finite-dimensional vector space. In the context of estimation, one popular and convenient notion of closeness between $\hat{\theta}$ and $\theta(\vartheta)$ is the distance in the Hilbert space $L^2(\Omega, \mathcal{A}, \mathbb{P}^\vartheta)$. This leads us to the following definition.

Definition 30.2. Given $\vartheta \in \Theta$, the corresponding *mean squared error* for an estimator $\hat{\theta}$ of a real-valued parameter θ is defined by

$$\mathrm{MSE}^\vartheta(\hat{\theta}, \theta) = E^\vartheta\big[\big(\hat{\theta} - \theta(\vartheta)\big)^2\big] \in [0, +\infty],$$

where E^ϑ denotes the expected value of a random variable with respect to the probability measure $\mathbb{P}^\vartheta$. In other words, $\mathrm{MSE}^\vartheta(\hat{\theta}, \theta)$ is the squared $L^2(\Omega, \mathcal{A}, \mathbb{P}^\vartheta)$-distance between $\hat{\theta}$ and the constant $\theta(\vartheta)$ (or $+\infty$ if $\hat{\theta}$ is not $\mathbb{P}^\vartheta$-square integrable).

Note that the mean squared error of $\hat{\theta}$ depends on the true value ϑ of the full parameter and thus one would typically want an estimator to have a small mean squared error for every $\vartheta \in \Theta$, as the true value of ϑ is unknown.

The mean squared error of an estimator is closely related to its variance. More explicitly, let $\vartheta \in \Theta$ be fixed and assume that the estimator $\hat{\theta}$ is $\mathbb{P}^\vartheta$-square integrable, so that its corresponding mean squared error is finite. Since $E^\vartheta(\hat{\theta})$ is the $L^2(\Omega, \mathcal{A}, \mathbb{P}^\vartheta)$-orthogonal projection of $\hat{\theta}$ onto the subspace of almost surely constant maps, we have that the two terms appearing in the sum on the righthand side of the equality

$$\hat{\theta} - \theta(\vartheta) = [\hat{\theta} - E^\vartheta(\hat{\theta})] + [E^\vartheta(\hat{\theta}) - \theta(\vartheta)]$$

are $L^2(\Omega, \mathcal{A}, \mathbb{P}^\vartheta)$-orthogonal and therefore the mean squared error of $\hat{\theta}$ can be written as:

$$\begin{aligned} \mathrm{MSE}^\vartheta(\hat{\theta}, \theta) &= E^\vartheta\big[\big(\hat{\theta} - E^\vartheta(\hat{\theta})\big)^2\big] + \big(E^\vartheta(\hat{\theta}) - \theta(\vartheta)\big)^2 \\ &= \mathrm{Var}^\vartheta(\hat{\theta}) + \big(E^\vartheta(\hat{\theta}) - \theta(\vartheta)\big)^2, \end{aligned} \tag{30.1}$$

where Var^ϑ denotes the variance of a random variable with respect to the probability measure $\mathbb{P}^\vartheta$.

Definition 30.3. Given $\vartheta \in \Theta$, the corresponding *bias* of a $\mathbb{P}^\vartheta$-integrable estimator $\hat{\theta}$ for a parameter θ is the difference

$$E^\vartheta(\hat{\theta}) - \theta(\vartheta)$$

between the expected value of the estimator and the true value of the parameter. An estimator which has zero bias for all $\vartheta \in \Theta$ is called *unbiased*.

Using this terminology, formula (30.1) can be nicely stated as follows.

Proposition 30.4. *Let the value ϑ of the full parameter of a stochastic model be fixed. If an estimator $\hat{\theta}$ for some real-valued parameter θ is integrable then the mean squared error of $\hat{\theta}$ is equal to the sum of the variance of $\hat{\theta}$ with the square of the bias of $\hat{\theta}$.*

Proof. If $\hat{\theta}$ is square integrable then this is just (30.1). Otherwise, both the variance and the mean squared error of $\hat{\theta}$ are infinite. □

Absence of bias is usually considered a desirable property for an estimator, but bias is not always as bad as it sounds. Note that if an estimator is unbiased then its mean squared error is equal to its variance and thus an unbiased estimator with a small variance is a good estimator. However, in some situations we might have an estimator that is biased but has a smaller mean squared error than some other estimator that is unbiased. In this case, the biased estimator might be preferable. Note also that if $f : \mathbb{R} \to \mathbb{R}$ is a continuous function then the equality $E^{\vartheta}\big(f(\hat{\theta})\big) = f\big(E^{\vartheta}(\hat{\theta})\big)$ typically does not hold, except for somewhat trivial cases (like the case in which f is affine). Thus, if $\hat{\theta}$ is an unbiased estimator for θ it will typically be false that $f(\hat{\theta})$ is an unbiased estimator for $f(\theta)$, although $f(\hat{\theta})$ is often used as an estimate for $f(\theta)$ when $\hat{\theta}$ is used as an estimate for θ.

Another relevant property of an estimator $\hat{\theta}$ is that it should have a small probability of assuming a value that is far away from the value of the parameter θ. The following simple inequality shows that this property will hold if $\hat{\theta}$ has a small mean squared error.

Proposition 30.5. *If X is a random variable and $c \in \mathbb{R}$ is a real number then for every $\varepsilon > 0$ the following inequality holds:*

$$\mathbb{P}(|X - c| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\, E\big((X - c)^2\big).$$

Proof. Simply note that:

$$E\big((X - c)^2\big) \geq \int_{[|X-c|\geq\varepsilon]} (X - c)^2 \,\mathrm{d}\mathbb{P} \geq \varepsilon^2\, \mathbb{P}(|X - c| \geq \varepsilon). \qquad \square$$

Corollary 30.6 (Chebyshev inequality)**.** *If X is a random variable with finite expectation then for every $\varepsilon > 0$ the following inequality holds:*

$$\mathbb{P}\big(|X - E(X)| \geq \varepsilon\big) \leq \frac{1}{\varepsilon^2}\, \mathrm{Var}(X).$$

Proof. Apply Proposition 30.5 with $c = E(X)$. □

Corollary 30.7. *Let $(\mathbb{P}_X^{\vartheta})_{\vartheta\in\Theta}$ be a stochastic model and $\theta : \Theta \to \mathbb{R}$ be a real-valued parameter. If $\hat{\theta}$ is an estimator for θ then for every $\varepsilon > 0$ the following inequality holds*

$$\mathbb{P}^{\vartheta}\big(|\hat{\theta} - \theta(\vartheta)| \geq \varepsilon\big) \leq \frac{1}{\varepsilon^2}\, \mathrm{MSE}^{\vartheta}(\hat{\theta}, \theta),$$

for every $\vartheta \in \Theta$. □

Example 30.8. Let $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ be a stochastic model such that $X = (X_i)_{i=1}^n$ is an $\mathbb{R}^n$-valued random vector. Moreover, assume that for every $\vartheta \in \Theta$ the family $(X_i)_{i=1}^n$ is i.i.d. and each X_i has finite expected value with respect to $\mathbb{P}^\vartheta$. Let $\mu : \Theta \to \mathbb{R}$ be the parameter defined by $\mu(\vartheta) = E^\vartheta(X_i)$, for all $\vartheta \in \Theta$ and any $i = 1, \ldots, n$. The parameter μ is called the *population mean*. The most commonly used estimator for μ is the *sample mean* defined by

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

We have

$$E^\vartheta(\bar{X}) = \frac{1}{n}\sum_{i=1}^n E^\vartheta(X_i) = \mu(\vartheta),$$

for all $\vartheta \in \Theta$ and therefore the estimator $\bar{X}$ is unbiased. Now assume that the variance $\mathrm{Var}^\vartheta(X_i) = \sigma^2(\vartheta)$ is finite for every $\vartheta \in \Theta$. The variance of $\bar{X}$ is then easily computed as

$$\mathrm{Var}^\vartheta(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}^\vartheta(X_i) = \frac{\sigma^2(\vartheta)}{n},$$

since the covariance between X_i and X_j is zero for $i \neq j$. Note that the variance of $\bar{X}$ coincides with its mean squared error as $\bar{X}$ is unbiased. Moreover, the mean squared error of $\bar{X}$ tends to zero as $n \to +\infty$ for every $\vartheta \in \Theta$.

The situation considered in Example 30.8 can be seen as a particular case of the following more general set up: the data $X = (X_i)_{i=1}^n$ is a simple random sample of some probability measure and the parameter that we want to estimate is a function of that probability measure. The strategy that we used to obtain an estimator for the parameter in Example 30.8 could be shortly described as "do to the sample the same thing that one does to the population to calculate the value of the parameter". By "population" we mean the probability measure $\mathbb{P}_{X_i}^\vartheta$ from which the sample is taken. In Example 30.8 what is done to the population to obtain the value of the parameter μ is computing the mean (i.e., the expected value of a random variable whose distribution is the probability measure from which the sample is taken). The estimator of μ is then obtained by doing the same thing to the sample, i.e., we compute the mean of the sample. In general, it is not clear what the recipe "do the same thing to the sample as what was done to the population" means, as the population is a probability measure and the observed value of the sample is an n-tuple. We now clarify the recipe by showing how to associate a probability measure to an n-tuple.

Definition 30.9. Let $(M, \mathcal{B})$ be a measurable space and $x = (x_i)_{i=1}^n$ be an n-tuple of elements of M. The *empirical distribution* associated to the

n-tuple x is the probability measure $\mathbb{P}_x^{\mathrm{emp}} : \mathcal{B} \to [0,1]$ defined by

$$\mathbb{P}_x^{\mathrm{emp}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i},$$

where $\delta_{x_i} : \mathcal{B} \to [0,1]$ denotes the Dirac delta probability measure corresponding to the point $x_i \in M$.

Note that $\mathbb{P}_x^{\mathrm{emp}}$ can be alternatively described as the push-forward under the map $\{1,\dots,n\} \ni i \mapsto x_i \in M$ of the discrete uniform distribution on $\{1,\dots,n\}$. We can thus think of $\mathbb{P}_x^{\mathrm{emp}}$ as the probability measure that models the experiment of choosing a term from the sequence $(x_i)_{i=1}^n$ in such a way that every term (more precisely, every index i) has the same probability of being chosen.

Clearly, if $f : M \to \mathbb{R}$ is a measurable function then:

$$\int_M f \,\mathrm{d}\mathbb{P}_x^{\mathrm{emp}} = \frac{1}{n}\sum_{i=1}^{n} f(x_i).$$

In particular, if $(M, \mathcal{B})$ is the real line endowed with the Borel σ-algebra then the expected value of a random variable with distribution $\mathbb{P}_x^{\mathrm{emp}}$ is simply the mean $\frac{1}{n}\sum_{i=1}^n x_i$. The recipe "do to the sample the same thing that one does to the population to calculate the value of the parameter" can then be precisely formulated as follows: do to the empirical distribution $\mathbb{P}_x^{\mathrm{emp}}$ the same thing that one does to the probability measure $\mathbb{P}_{X_i}^{\vartheta}$ to calculate the value of the parameter, where $x = (x_i)_{i=1}^n$ is the observed value of the random sample $(X_i)_{i=1}^n$.

Let us see what happens if we apply this recipe to the variance of the population instead of the mean.

Example 30.10. Let $(\mathbb{P}_X^{\vartheta})_{\vartheta\in\Theta}$ be a stochastic model such that $X = (X_i)_{i=1}^n$ is an $\mathbb{R}^n$-valued random vector. Moreover, assume that for every $\vartheta \in \Theta$ the family $(X_i)_{i=1}^n$ is i.i.d. and each X_i has finite variance with respect to $\mathbb{P}^{\vartheta}$. Let $\sigma : \Theta \to \mathbb{R}$ be the parameter defined by $\sigma(\vartheta) = [\mathrm{Var}^{\vartheta}(X_i)]^{\frac{1}{2}}$, for all $\vartheta \in \Theta$ and any $i = 1,\dots,n$. The parameter σ^2 is called the *population variance*. Given a sequence $x = (x_i)_{i=1}^n \in \mathbb{R}^n$, the variance of a random variable whose distribution is the empirical distribution $\mathbb{P}_x^{\mathrm{emp}}$ is given by:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \Big(\frac{1}{n}\sum_{i=1}^{n} x_i^2\Big) - \bar{x}^2,$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. The "do to the sample what you do to the population" recipe says then that we should use

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{30.2}$$

as an estimator of σ^2, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ as in Example 30.8. Let us compute the expected value of the estimator (30.2). This can be easily done

by brute force, but better insights are obtained if we use instead the results of Section 25.

Let $\mathbb{R}^n$ be endowed with its canonical inner product and denote by $\|\cdot\|$ the corresponding norm. Note that $(\bar{x},\dots,\bar{x}) \in \mathbb{R}^n$ is the orthogonal projection of $x \in \mathbb{R}^n$ onto the one-dimensional subspace of $\mathbb{R}^n$ consisting of constant n-tuples. Thus, if $P : \mathbb{R}^n \to \mathbb{R}^n$ denotes the orthogonal projection onto the orthogonal complement of that one-dimensional subspace we have

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \|P(x)\|^2$$

and therefore (30.2) is equal to $\frac{1}{n}\|P(X)\|^2$. Let us compute the expected value of $\|P(X)\|^2$ using Corollary 25.3. The variance Σ of the random vector X with respect to the probability measure $\mathbb{P}^\vartheta$ is equal to $\sigma^2(\vartheta)$ times the canonical inner product of $\mathbb{R}^{n*}$ and, assuming that $\sigma(\vartheta) \neq 0$, we have that Σ^{-1} is equal to $\frac{1}{\sigma^2(\vartheta)}$ times the canonical inner product of $\mathbb{R}^n$. Of course, P is also an orthogonal projection with respect to the inner product Σ^{-1} and the norm associated to Σ^{-1} is $\frac{1}{\sigma(\vartheta)}\|\cdot\|$. Since all coordinates of X have the same expected value we have $P\big(E^\vartheta(X)\big) = 0$ and therefore Corollary 25.3 yields:

$$\frac{1}{\sigma^2(\vartheta)}\,E^\vartheta\big(\|P(X)\|^2\big) = n-1. \tag{30.3}$$

We conclude that the expected value of (30.2) is equal to

$$E^\vartheta\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\Big) = \frac{1}{n}\,E^\vartheta\big(\|P(X)\|^2\big) = \frac{n-1}{n}\,\sigma^2(\vartheta). \tag{30.4}$$

Note that formula (30.4) holds trivially if $\sigma(\vartheta) = 0$ as in that case $X_i - \bar{X} = 0$ almost surely.

We have established that the estimator (30.2) obtained by using the "do to the sample what you do to the population" recipe is biased! From (30.3) we readily see that an unbiased estimator for σ^2 is given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2,$$

if $n \geq 2$. The estimator S^2 is known as the *sample variance*. Thus, what we call "sample variance" does not coincide with the variance of the sample, i.e., the variance of the empirical distribution associated to the observed value of the sample. The reason why we do not define the sample variance as simply the variance of the sample is because the variance of the sample is a biased estimator of the variance of the population.

We now wish to compute the variance of the estimators S^2 and (30.2). We make a digression to develop a little theory in order to avoid doing the computation by brute force.

Definition 30.11. Let V and W be real finite-dimensional vector spaces endowed with inner products $\langle\cdot,\cdot\rangle_V$ and $\langle\cdot,\cdot\rangle_W$, respectively. The *Hilbert–Schmidt inner product* on the space of linear transformations from V to W is defined by

$$\langle T, S\rangle_{\mathrm{HS}} = \mathrm{tr}(T^{\mathrm{t}} \circ S),$$

for every pair of linear transformations $T : V \to W$ and $S : V \to W$, where $T^{\mathrm{t}} : W \to V$ denotes the *transpose* of T with respect to the given inner products, i.e., T^{t} is characterized by the equality

$$\langle T(v), w\rangle_W = \langle v, T^{\mathrm{t}}(w)\rangle_V,$$

for all $v \in V$ and $w \in W$. The norm $\|\cdot\|_{\mathrm{HS}}$ associated to the Hilbert–Schmidt inner product is called the *Hilbert–Schmidt norm*.

Clearly, if T and S are represented by matrices $(T_{ij})_{n\times m}$ and $(S_{ij})_{n\times m}$ with respect to orthonormal bases we have $\langle T, S\rangle_{\mathrm{HS}} = \sum_{i=1}^{n}\sum_{j=1}^{m} T_{ij}S_{ij}$. Moreover, if $P : V \to V$ is an orthogonal projection onto some subspace of V then $\|P\|^2_{\mathrm{HS}} = \langle P, P\rangle_{\mathrm{HS}}$ is equal to the dimension of the image of P.

Definition 30.12. Let X be a random variable whose fourth power X^4 is integrable and assume that X is not almost surely constant. The *kurtosis* of X is defined by

$$\mathrm{Kurt}(X) = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^4\Big],$$

where $\mu = E(X)$ and $\sigma = \mathrm{Var}(X)^{\frac{1}{2}}$.

The expression $\frac{X-\mu}{\sigma}$ is called the *standardized version* of X and it remains unchanged if we add a constant to X or multiply X by a nonzero constant. In particular, the kurtosis of X is invariant by such operations. Note that using the equality

$$E(Y^2) = \mathrm{Var}(Y) + E(Y)^2$$

with $Y = \big(\frac{X-\mu}{\sigma}\big)^2$ we obtain

$$\mathrm{Kurt}(X) = \mathrm{Var}\left[\Big(\frac{X-\mu}{\sigma}\Big)^2\right] + 1$$

and in particular $\mathrm{Kurt}(X) \geq 1$. The minimum $\mathrm{Kurt}(X) = 1$ is attained if and only if $(X-\mu)^2$ is almost surely constant and it is not hard to check that this happens if and only if

$$\mathbb{P}(X = x_0) = \mathbb{P}(X = x_1) = \frac{1}{2},$$

for two distinct real numbers x_0 and x_1.

If X is a nondegenerate normal random variable then the kurtosis of X is equal to the kurtosis of a standard normal random variable and is therefore given by:

$$\mathrm{Kurt}(X) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} x^4 e^{-\frac{x^2}{2}}\,\mathrm{d}\mathfrak{m}(x) = 3;$$

namely, the integral in the formula above is easily computed by writing the integrand as $x^3(xe^{-\frac{x^2}{2}})$ and using integration by parts to show that the integral is equal to 3 times the variance of a standard normal random variable. The difference

$$\mathrm{Kurt}(X) - 3$$

between the kurtosis of X and the kurtosis of a nondegenerate normal random variable is known as the *excess kurtosis* of X. A random variable with a high kurtosis can be informally described as a random variable whose distribution has *heavy tails*.

We now use kurtosis to obtain a formula for the expected value of the square of a quadratic form of an i.i.d. finite family of random variables with null expected values.

Lemma 30.13. *Let $(X_i)_{i=1}^n$ be an independent n-tuple of random variables and assume that $E(X_i) = 0$, $\mathrm{Var}(X_i) = \sigma^2$ and $\mathrm{Kurt}(X_i) = \kappa$, for all $i = 1, \ldots, n$, for certain $\sigma > 0$ and $\kappa \geq 1$. If $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation represented with respect to the canonical basis by a matrix $(a_{ij})_{n\times n}$ and if $\langle \cdot, \cdot \rangle$ denotes the canonical inner product of $\mathbb{R}^n$ then:*

$$E\big(\langle T(X), X\rangle^2\big) = \sigma^4\Big([\mathrm{tr}(T)]^2 + \frac{1}{2}\|T + T^{\mathrm{t}}\|_{\mathrm{HS}}^2 + (\kappa - 3)\sum_{i=1}^n (a_{ii})^2\Big).$$

Proof. We have $\langle T(X), X\rangle = \sum_{i=1}^n \sum_{j=1}^n a_{ij} X_i X_j$ and:

$$\langle T(X), X\rangle^2 = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n a_{ij} a_{kl} X_i X_j X_k X_l. \tag{30.5}$$

Note that if $i \notin \{j, k, l\}$ then $E(X_i X_j X_k X_l) = E(X_i)E(X_j X_k X_l) = 0$. Similarly, if any element of $\{1, \ldots, n\}$ occurs exactly once in the sequence (i, j, k, l) then $E(X_i X_j X_k X_l) = 0$. It follows that the expected value of the sum in (30.5) is given by:

$$\sum_{\substack{i,j=1 \\ i\neq j}}^n a_{ii} a_{jj} \sigma^4 + \sum_{\substack{i,j=1 \\ i\neq j}}^n (a_{ij})^2 \sigma^4 + \sum_{\substack{i,j=1 \\ i\neq j}}^n a_{ij} a_{ji} \sigma^4 + \sum_{i=1}^n (a_{ii})^2 \kappa \sigma^4.$$

To conclude the proof note that

$$\sum_{\substack{i,j=1 \\ i\neq j}}^n (a_{ij})^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^n a_{ij} a_{ji} = \frac{1}{2} \sum_{\substack{i,j=1 \\ i\neq j}}^n (a_{ij} + a_{ji})^2 = \frac{1}{2}\|T + T^{\mathrm{t}}\|_{\mathrm{HS}}^2 - 2\sum_{i=1}^n (a_{ii})^2$$

and:

$$\sum_{\substack{i,j=1 \\ i\neq j}}^n a_{ii} a_{jj} = [\mathrm{tr}(T)]^2 - \sum_{i=1}^n (a_{ii})^2. \qquad \square$$

Corollary 30.14. *Let $(X_i)_{i=1}^n$ be an independent n-tuple of random variables and assume that $E(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$ and $\mathrm{Kurt}(X_i) = \kappa$, for all $i = 1, \dots, n$, for certain $\mu \in \mathbb{R}$, $\sigma > 0$ and $\kappa \geq 1$. We have:*

$$E\Big[\Big(\sum_{i=1}^n (X_i - \bar{X})^2\Big)^2\Big] = \sigma^4(n-1)\Big[n + 1 + (\kappa - 3)\Big(1 - \frac{1}{n}\Big)\Big],$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. In particular:

$$\mathrm{Var}\Big(\sum_{i=1}^n (X_i - \bar{X})^2\Big) = \sigma^4(n-1)\Big[2 + (\kappa - 3)\Big(1 - \frac{1}{n}\Big)\Big].$$

Proof. By replacing X_i with $X_i - \mu$ we can assume without loss of generality that $E(X_i) = 0$, for all $i = 1, \dots, n$. If $P : \mathbb{R}^n \to \mathbb{R}^n$ denotes the orthogonal projection onto the orthogonal complement of $(1, \dots, 1)$ then:

$$\langle P(X), X\rangle = \langle P(X), P(X)\rangle = \sum_{i=1}^n (X_i - \bar{X})^2. \tag{30.6}$$

The formula for the expected value of the square of (30.6) follows from Lemma 30.13 by noting that $\mathrm{tr}(P) = n - 1$, $P = P^{\mathrm{t}}$, $\|P\|_{\mathrm{HS}}^2 = n - 1$ and that the diagonal elements of the matrix that represents P with respect to the canonical basis of $\mathbb{R}^n$ are all equal to $1 - \frac{1}{n}$. The formula for the variance of (30.6) then follows from the fact that the expected value of (30.6) is equal to $(n-1)\sigma^2$ (recall (30.3)). □

Example 30.15. Going back to the set up of Example 30.10, assume in addition that every X_i has an integrable fourth power and that it is not almost surely constant. Denoting by $\kappa(\vartheta)$ the kurtosis of X_i with respect to $\mathbb{P}^\vartheta$, Corollary 30.14 gives us:

$$\mathrm{Var}(S^2) = \frac{\sigma^4(\vartheta)}{n-1}\Big[2 + \big(\kappa(\vartheta) - 3\big)\Big(1 - \frac{1}{n}\Big)\Big] = \frac{\sigma^4(\vartheta)}{n}\Big(\frac{2n}{n-1} + \kappa(\vartheta) - 3\Big).$$

Note that the variance of S^2 tends to zero for fixed ϑ as n tends to $+\infty$ and that, since the estimator S^2 is unbiased, its variance coincides with its mean squared error. Let us now compute the mean squared error of the biased estimator $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ and, more generally, the mean squared error of an estimator for σ^2 of the form

$$\frac{1}{\alpha}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{\alpha}\, S^2,$$

where α is a positive constant. Since the mean squared error is the sum of the variance with the square of the bias and since the bias of $\frac{n-1}{\alpha}S^2$ is equal to $\big(\frac{n-1}{\alpha} - 1\big)\sigma^2(\vartheta)$, we obtain

$$\mathrm{MSE}^\vartheta\Big(\frac{n-1}{\alpha}\, S^2, \sigma^2\Big) = \frac{(n-1)^2}{\alpha^2}\,\mathrm{Var}(S^2) + \Big(\frac{n-1}{\alpha} - 1\Big)^2\sigma^4(\vartheta)$$

and a straightforward computation yields:

$$\mathrm{MSE}^{\vartheta}\Big(\frac{1}{\alpha}\sum_{i=1}^{n}(X_i-\bar{X})^2,\sigma^2\Big)=\frac{n-1}{\alpha^2}\sigma^4(\vartheta)\Big(n+1+\frac{n-1}{n}\big(\kappa(\vartheta)-3\big)\Big)$$
$$-2\sigma^4(\vartheta)\frac{n-1}{\alpha}+\sigma^4(\vartheta).$$

This expression is a second degree polynomial in $\frac{1}{\alpha}$ and using the fact that $\kappa(\vartheta)\geq 1$ we see that its leading coefficient is positive. One then easily verifies that the minimum value of the mean squared error is attained at:

$$\alpha=n+1+\frac{n-1}{n}\big(\kappa(\vartheta)-3\big).$$

This is always larger than $n-1$ and therefore the unbiased estimator S^2 never coincides with the estimator of the form $\frac{1}{\alpha}\sum_{i=1}^{n}(X_i-\bar{X})^2$ that minimizes the mean squared error. In general, we cannot determine the optimal value of α as it depends on the kurtosis $\kappa(\vartheta)$ which depends on the unknown parameter ϑ. However, if the variables X_i are normal then $\kappa(\vartheta)=3$ for all $\vartheta\in\Theta$ and the optimum value is $\alpha=n+1$.

30.1. **Asymptotic theory of estimators.** Now let us study the behaviour of an estimator when the sample size goes to infinity. Strictly speaking, the previous sentence is not correctly formulated since a estimator is a specific function of a specific set of data X and therefore it is associated to a specific sample size. So, for example, the sample mean $\bar{X}$ discussed in Example 30.8 is defined as $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ and it is the mean of a sample of size n. If we change the value of n, we get a new estimator. Thus, the rigorous formulation of what we intend to do in this subsection is to study the behaviour of a *sequence* of estimators (indexed by sample size) when the sample size goes to infinity.

In order to formulate certain limit properties of a sequence of estimators, we need all of them to be defined on the same probability space and to achieve that goal we have to consider a model in which the data X is an infinite sample. For example, consider a stochastic model $(\mathbb{P}_X^{\vartheta})_{\vartheta\in\Theta}$ such that X takes values in the space of infinite sequences of real numbers (endowed with the product of the Borel σ-algebras) and $X=(X_i)_{i\geq 1}$ is an infinite sequence of random variables which is i.i.d. with respect to $\mathbb{P}^{\vartheta}$, for all $\vartheta\in\Theta$. Moreover, assume that $\mu(\vartheta)=E^{\vartheta}(X_i)$ is finite, for all $\vartheta\in\Theta$. For each $n\geq 1$, we consider the random variable

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

which we call the *sample mean for a sample of size* n. The distribution of $\bar{X}_n$ only depends on the distribution of the finite sequence $(X_i)_{i=1}^n$ and therefore the expected value and variance of $\bar{X}_n$ are the same as those computed in Example 30.8, i.e., the fact that the data X contains more than just $(X_i)_{i=1}^n$

does not alter the results of what we have already computed. In particular, $\bar{X}_n$ is an unbiased estimator of μ and, if X_i has finite variance, then the variance of $\bar{X}_n$ tends to zero as n tends to infinity. What is different from Example 30.8 is that now all of the random variables $\bar{X}_n$ are defined on the same probability space, which is a more suitable set up for studying limit properties of $(\bar{X}_n)_{n\geq 1}$ as n tends to infinity.

Definition 30.16. Let $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ be a stochastic model and let $\theta : \Theta \to \Theta'$ be a parameter for the model such that the set Θ' is endowed with a separable[5] metric and the corresponding Borel σ-algebra. A sequence of estimators $(\hat{\theta}_n)_{n\geq 1}$ for θ is said to be *consistent* (sometimes also called *weakly consistent*) if for every $\vartheta \in \Theta$ the sequence of random objects $(\hat{\theta}_n)_{n\geq 1}$ converges in probability to the constant random object $\theta(\vartheta)$ with respect to the probability measure $\mathbb{P}^\vartheta$. The sequence $(\hat{\theta}_n)_{n\geq 1}$ is said to be *strongly consistent* if for every $\vartheta \in \Theta$ the sequence of random objects $(\hat{\theta}_n)_{n\geq 1}$ converges almost surely to the constant $\theta(\vartheta)$ with respect to the probability measure $\mathbb{P}^\vartheta$.

Informally, people will usually say that "a certain estimator is consistent" or that "a certain estimator is strongly consistent", instead of attributing the property of consistency to a sequence of estimators. That is because we normally think of an estimator as corresponding to a certain strategy for computing an estimate of the parameter and such strategy can be naturally adapted to arbitrary sample sizes. However, as discussed at the beginning of the subsection, a single estimator corresponds to a single sample size, so the formally correct way of talking about consistency requires a sequence of estimators.

A simple way of establishing the (weak) consistency of a sequence of estimators is to show that the limit of their mean squared errors is zero.

Proposition 30.17. *Let $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ be a stochastic model, $\theta : \Theta \to \mathbb{R}$ be a real-valued parameter for the model and $(\hat{\theta}_n)_{n\geq 1}$ be a sequence of estimators for θ. If*

$$\lim_{n\to+\infty} \mathrm{MSE}^\vartheta(\hat{\theta}_n, \theta) = 0, \tag{30.7}$$

for all $\vartheta \in \Theta$ then the sequence $(\hat{\theta}_n)_{n\geq 1}$ is consistent.

Proof. Condition (30.7) simply says that the sequence $(\hat{\theta}_n)_{n\geq 1}$ converges in $L^2(\Omega, \mathcal{A}, \mathbb{P}^\vartheta)$ to the constant $\theta(\vartheta)$ and thus the result is simply a restatement of the simple fact that a sequence that converges in L^2 converges in probability. It is also a direct consequence of Corollary 30.7. □

We observe that both the notion of mean squared error and Proposition 30.17 can be readily generalized to the case when θ takes values in an arbitrary separable metric space.

[5]See Section 12 for an explanation of why we only study convergence in probability for sequences of random objects taking values in a separable metric space.

Corollary 30.18. *Let $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ be a stochastic model such that X takes values in the space of infinite sequences of real numbers. Moreover, assume that for every $\vartheta \in \Theta$, $(X_i)_{i\geq 1}$ is an infinite sequence of random variables which is i.i.d. with respect to $\mathbb{P}^\vartheta$ and such that $E^\vartheta(X_i^2)$ is finite. If*

$$\mu(\vartheta) = E^\vartheta(X_i)$$

denotes the population mean and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ then $(\bar{X}_n)_{n\geq 1}$ is a consistent sequence of estimators for μ.

Proof. Follows from Proposition 30.17 using the results presented in Example 30.8. □

Corollary 30.19. *Let $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ be a stochastic model such that X takes values in the space of infinite sequences of real numbers. Moreover, assume that for every $\vartheta \in \Theta$, $(X_i)_{i\geq 1}$ is an infinite sequence of random variables which is i.i.d. with respect to $\mathbb{P}^\vartheta$ and such that $E^\vartheta(X_i^4)$ is finite. If*

$$\sigma^2(\vartheta) = \mathrm{Var}^\vartheta(X_i)$$

denotes the population variance and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ then $(S_n^2)_{n\geq 2}$ is a consistent sequence of estimators for σ^2.

Proof. Follows from Proposition 30.17 using the results presented in Examples 30.10 and 30.15. □

Note that the sequence $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n-1}{n} S_n^2$ of biased estimators of σ^2 is also consistent since $\lim_{n\to+\infty} \frac{n-1}{n} = 1$.

Remark 30.20. The celebrated *strong law of large numbers* (see [1, 6.2.5]) states that if $(X_i)_{i\geq 1}$ is an infinite i.i.d. sequence of random variables with finite expected value $E(X_i) = \mu$ then the sequence $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ converges almost surely to μ. This means that in Corollary 30.18 one can actually obtain the stronger thesis that $(\bar{X}_n)_{n\geq 1}$ is a strongly consistent estimator of μ under the weaker assumption that X_i has finite expected value with respect to $\mathbb{P}^\vartheta$, for all $\vartheta \in \Theta$. Moreover, noting that

$$S_n^2 = \frac{n}{n-1}\Big(\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2\Big)$$

and applying the strong law of large numbers to the sequence $(X_i^2)_{i\geq 1}$ we conclude that in Corollary 30.19 we can obtain the stronger thesis that $(S_n^2)_{n\geq 2}$ is a strongly consistent estimator for σ^2 under the weaker assumption that X_i is square integrable with respect to $\mathbb{P}^\vartheta$, for all $\vartheta \in \Theta$.

30.2. **Maximum likelihood estimation.** Let $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ be a stochastic model with the data X taking values in a measurable space $(M, \mathcal{B})$. Assume that there exists a σ-finite nonnegative countably additive measure

μ on $\mathcal{B}$ such that $\mathbb{P}_X^\vartheta$ is absolutely continuous with respect to μ for every $\vartheta \in \Theta$. We can then define a map $L : M \times \Theta \to [0, +\infty[$ such that

$$L(\cdot, \vartheta) = \frac{\mathrm{d}\mathbb{P}_X^\vartheta}{\mathrm{d}\mu} \tag{30.8}$$

for all $\vartheta \in \Theta$, i.e., such that $M \ni x \mapsto L(x, \vartheta) \in [0, +\infty[$ is a Radon–Nikodym derivative of $\mathbb{P}_X^\vartheta$ with respect to μ for every $\vartheta \in \Theta$. In other words, $L(\cdot, \vartheta)$ is a probability density function for X with respect to μ when the domain of X is endowed with the probability measure $\mathbb{P}^\vartheta$. If $x \in M$ is the observed value of the data X in some experiment then the function

$$\Theta \ni \vartheta \longmapsto L(x, \vartheta) \in [0, +\infty[$$

is known as the *likelihood function* associated to x. If such function attains its maximum at some point $\vartheta \in \Theta$ then such value of the full parameter can be seen as the value that is most compatible with the observed data x since, in some sense, it makes the value $X = x$ "more probable" than other values of the parameter. We note that in many important examples the actual probability $\mathbb{P}^\vartheta(X = x)$ of observing $X = x$ under ϑ will be zero for all $x \in M$ and all $\vartheta \in \Theta$ and that is why we have to work with probability densities instead of actual probabilities.

The value of ϑ that maximizes the likelihood function for a certain value $x \in M$ of the data is known as the *maximum likelihood estimate* of the full parameter corresponding to x. Let us write this as a formal definition.

Definition 30.21. Let $(\mathbb{P}_X^\vartheta)_{\vartheta \in \Theta}$ be a stochastic model with the data X taking values in a measurable space $(M, \mathcal{B})$. Assume that μ is a σ-finite nonnegative countably additive measure μ on $\mathcal{B}$ such that $\mathbb{P}_X^\vartheta$ is absolutely continuous with respect to μ for every $\vartheta \in \Theta$ and let $L : M \times \Theta \to [0, +\infty[$ be a function such that (30.8) holds. Assume that for every $x \in M$ the function $L(x, \cdot)$ attains its maximum at some point of Θ and let $f : M \to \Theta$ be a map such that $f(x)$ is a point of maximum of $L(x, \cdot)$, for every $x \in M$. If Θ is endowed with a σ-algebra and the map f is measurable then $\hat{\theta} = f(X)$ is called a *maximum likelihood estimator* of the full parameter $\theta : \Theta \to \Theta$ corresponding to the map L.

There is a rich asymptotic theory for maximum likelihood estimators [5, 7.3.2—7.3.5] which we will not develop here. We will finish the section with a few technical comments and a simple concrete example of maximum likelihood estimators.

There are two worries that immediately arise when considering maximum likelihood estimation. First of all, Radon–Nikodym derivatives are only unique up to μ-almost everywhere equality and thus evaluating the probability density function at a specific point $x \in M$ is meaningless, unless x happens to be a point with positive measure. However, as in Subsection 28.1, this uniqueness problem can be solved by adding a continuity requirement. More explicitly, if M is endowed with a topology and $\mathcal{B}$ is the

Borel σ-algebra, then a continuous probability density function with respect to the measure μ (if it exists) is uniquely defined at the points in the support of μ. A second concern is related to the dependence of maximum likelihood estimators on the choice of the measure μ. The following result deals with that.

Proposition 30.22. *Let $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ be a stochastic model with the data X taking values in a measurable space $(M,\mathcal{B})$. Let μ and ν be σ-finite nonnegative countably additive measures on $\mathcal{B}$ such that $\mathbb{P}_X^\vartheta$ is absolutely continuous with respect to both μ and ν, for all $\vartheta\in\Theta$. If $L : M\times\Theta\to[0,+\infty[$ is a function such that*

$$L(\cdot,\vartheta)=\frac{\mathrm{d}\mathbb{P}_X^\vartheta}{\mathrm{d}\mu} \tag{30.9}$$

holds for all $\vartheta\in\Theta$ then there exists a measurable function $g : M\to[0,+\infty[$ such that

$$L(\cdot,\vartheta)g=\frac{\mathrm{d}\mathbb{P}_X^\vartheta}{\mathrm{d}\nu}$$

holds, for all $\vartheta\in\Theta$.

Proof. By Lebesgue's Decomposition Theorem we can write $\mu=\mu_{\mathrm{ac}}+\mu_{\mathrm{s}}$, with μ_{ac} and μ_{s} σ-finite nonnegative countably additive measures on $\mathcal{B}$ such that μ_{ac} is absolutely continuous with respect to ν and μ_{s} and ν are mutually singular, i.e., M can be written as a disjoint union of $M_1\in\mathcal{B}$ and $M_2\in\mathcal{B}$ in such a way that $\mu_{\mathrm{s}}(M_2)=0$ and $\nu(M_1)=0$. For every $\vartheta\in\Theta$ we have

$$\begin{aligned}0\le\int_M L(x,\vartheta)\,\mathrm{d}\mu_{\mathrm{s}}(x)&=\int_{M_1}L(x,\vartheta)\,\mathrm{d}\mu_{\mathrm{s}}(x)\\&\le\int_{M_1}L(x,\vartheta)\,\mathrm{d}\mu(x)=\mathbb{P}_X^\vartheta(M_1)=0,\end{aligned}$$

because $\mathbb{P}_X^\vartheta$ is absolutely continuous with respect to ν. Thus $L(x,\vartheta)=0$ for μ_s-almost every $x\in M$ and this implies that $\mathbb{P}_X^\vartheta$ is absolutely continuous with respect to μ_{ac} and

$$L(\cdot,\vartheta)=\frac{\mathrm{d}\mathbb{P}_X^\vartheta}{\mathrm{d}\mu_{\mathrm{ac}}},$$

for all $\vartheta\in\Theta$. Hence the desired map $g : M\to[0,+\infty[$ can be taken as a Radon–Nikodym derivative of μ_{ac} with respect to ν. □

According to Proposition 30.22, if we replace the measure μ with another measure ν such that all $\mathbb{P}_X^\vartheta$ are absolutely continuous with respect to ν then the map L satisfying (30.9) can be replaced with $(x,\vartheta)\mapsto L(x,\vartheta)g(x)$, for some nonnegative measurable function g. If the observed data $x\in M$ satisfies $g(x)>0$, then both likelihoods $L(x,\cdot)$ and $L(x,\cdot)g(x)$ have the

same maximum points. Moreover, the set $g^{-1}(0)$ of bad data points has probability zero with respect to all $\mathbb{P}_X^\vartheta$, since

$$\mathbb{P}_X^\vartheta\big(g^{-1}(0)\big) = \int_{g^{-1}(0)} L(x,\vartheta)g(x)\,\mathrm{d}\nu = 0,$$

for all $\vartheta \in \Theta$.

Let us now discuss an elementary concrete example of maximum likelihood estimation involving a simple random sample from a normal distribution.

Example 30.23. Let $(\mathbb{P}_X^\vartheta)_{\vartheta\in\Theta}$ be a stochastic model such that X is an $\mathbb{R}^n$-valued random vector with $X = (X_i)_{i=1}^n$ i.i.d., $X_i \sim N(\mu,\sigma^2)$ with respect to $\mathbb{P}^\vartheta$ for all $\vartheta = (\mu,\sigma) \in \Theta$, where $\Theta = \mathbb{R}\times\,]0,+\infty[$. Let $L : \mathbb{R}\times\Theta \to [0,+\infty[$ be such that $L(\cdot,\vartheta)$ is a continuous Radon–Nikodym derivative of $\mathbb{P}_X^\vartheta$ with respect to Lebesgue measure for all $\vartheta\in\Theta$, so that L is given by

$$L(x,\mu,\sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\prod_{i=1}^n e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

for all $x\in\mathbb{R}^n$, $\mu\in\mathbb{R}$ and $\sigma>0$. For each $x\in\mathbb{R}^n$ we wish to determine the maximum points of the likelihood $(\mu,\sigma)\mapsto L(x,\mu,\sigma)$. It is a little easier to work with the logarithm of L (which of course has the same maximum points):

$$\ell(x,\mu,\sigma) = \ln[L(x,\mu,\sigma)] = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}.$$

The function $(\mu,\sigma)\mapsto\ell(x,\mu,\sigma)$ is known as the *log-likelihood*. For fixed $\sigma>0$, $\ell(x,\mu,\sigma)$ attains its unique global maximum when $\mu\in\mathbb{R}$ is such that $(\mu,\ldots,\mu)$ is closest to x with respect to standard Euclidean distance. Therefore, such maximum is attained at $\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, which is the value of μ that makes $(\mu,\ldots,\mu)$ equal to the orthogonal projection of x onto the subspace generated by $(1,\ldots,1)$. Now we must find $\sigma>0$ that maximizes $\ell(x,\bar{x},\sigma)$. By studying the sign of the derivative of $\ell(x,\bar{x},\sigma)$ with respect to σ we see that $\ell(x,\bar{x},\sigma)$ attains its unique global maximum at

$$\sigma = \Big(\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2\Big)^{\frac{1}{2}},$$

provided that x_1, …, x_n are not all equal, i.e., provided that x does not belong to the subspace generated by $(1,\ldots,1)$. Hence, assuming $n\geq 2$ and removing the subspace generated by $(1,\ldots,1)$ from the counter-domain of X (which has probability zero with respect to $\mathbb{P}_X^\vartheta$, for all ϑ), we obtain that

$$\Big(\bar{X}, \Big(\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2\Big)^{\frac{1}{2}}\Big)$$

is a maximum likelihood estimator for (μ, σ), where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Note that the estimator that we have obtained for μ is just the unbiased estimator $\bar{X}$ discussed in Example 30.8, while the estimator for σ^2 is the biased estimator $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and not the unbiased estimator S^2 discussed in Example 30.10.

31. Hypothesis testing

References

[1] R. Ash & C. A. Doléans-Dade, *Probability & Measure Theory*, Second Edition, Academic Pres, 2000.

[2] H. Federer, *Geometric Measure Theory*, Springer–Verlag, 1969.

[3] A. Gelman & D. Nolan, *You can load a die, but you can't bias a coin*, The American Statistician, 56:4 (2002), pgs. 308—311, DOI: 10.1198/000313002605

[4] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, Third Edition, 1987.

[5] M. J. Schervish, *Theory of Statistics*, Springer–Verlag, 1995.

[6] S. M. Srivastava, *A course on Borel sets*, Springer–Verlag, 1998.

[7] D. V. Tausk, *Variables*, https://www.ime.usp.br/~tausk/texts/Variables.pdf, 2019.

Departamento de Matemática,
Universidade de São Paulo, Brazil
Email address: tausk@ime.usp.br
URL: http://www.ime.usp.br/~tausk