Introduction to Probability Theory for Economists

Enrico Scalas[1, *]

[1] *Laboratory on Complex Systems. Dipartimento di Scienze e Tecnologie Avanzate,*
Università del Piemonte Orientale "Amedeo Avogadro",
Via Bellini 25 G, 15100 Alessandria, Italy

(Dated: October 22, 2008)

Abstract

These lecture notes contain an introduction to the elementary theory of probability. The following topics will be presented:

- The meaning of probability;
- Kolmogorov's axioms;
- Random variables;
- Introduction to stochastic processes;
- Markov chains;
- Poisson process;
- Wiener process.
- Important distributions;

These notes contain further pointers to the literature as well as a selection of solved exercises.

PACS numbers: 02.50.-r, 02.50.Ey, 05.40.-a, 05.40.Jc,

[*] Electronic address: `enrico.scalas@mfn.unipmn.it`; URL: `www.mfn.unipmn.it/~scalas`

I. THE MEANING OF PROBABILITY

This section contains a short outline of the history of probability and a brief account of the debate on the meaning of probability. The two issues are interwoven. After reading and studying this section you should be able to

- present the history of probability;
- understand the main interpretations of probability (classical, frequentist, subjectivist, logical empiricist);
- compute probabilities of events based on the fundamental counting principle and combinatorial formulae.
- relate the interpretations to the history of human thought (especially if you already know something about philosophy).
- discuss some of the early applications of probability to Economics.

A. Early accounts and the birth of mathematical probability

We know that, in the 17th century, probability theory begun with the analysis of games of chance (a.k.a gambling). However, dice were already in use in ancient civilizations. Just to limit ourselves to the Mediterranean area (due to the somewhat Eurocentric culture of this author), dice are found in archaeological sites in Egypt. According to Svetonius (a Roman historian), in the first century, Emperor Claudius wrote a book on gambling, but unfortunately nothing of his book remains nowadays.

It is however true that chance has been a part of the life of our ancestors. Always (and this is true also today), individuals and societies have been faced with unpredictable events and it is not surprising that this unpredictability has been the subject of many discussions and speculations especially when compared with better predictable events such as the astronomical ones.

It is perhaps harder to understand why there had been no mathematical formalizations of probability theory until the 17th century. There is a remote possibility that, in ancient times, other treatises like the one of Claudius were available and also were lost, but, if this was the case, mathematical probability was perhaps a fringe subject.

A natural place for the development of probabilistic thought could have been Physics. Real measurements are never exactly in agreement with theory, and the measured values often fluctuate. Outcomes of physical experiments are a natural candidate for the development of a theory of random variables. However, the early developments of Physics did not take chance into account and the tradition of Physics remains far from probability. Even today, education in Physics virtually neglects probability theory.

The birth of probability theory was triggered by a letter sent to Blaise Pascal by his friend Chevalier De Méré on a dice gambling problem. In the 17th century in Europe, there were people rich enough to travel along the continent and waste their money gambling. De Méré was one of them. In the summer of 1654, Pascal wrote to Pierre de Fermat in order to solve De Méré's problem and out of their correspondence mathematical probability theory was born.

B. Laplace and the classical definition of probability

In 1812, Pierre Simon de Laplace published his celebrated book *Théorie analytique des probabilités* where he gave and discussed a definition of probability. In the meantime, Christian Huygens had written a first exhaustive book on mathematical probability, based on the ideas of Pascal and Fermat, *De ratiociniis in ludo aleae*, published in 1657 and the theory had been further developed by Jacob Bernoulli in *Ars conjectandi* published in 1713, seven years after his death. It is indeed in the 18th century, namely in 1733, that Daniel Bernoulli published the first paper where probability theory is applied to economics, *Specimen theoriae novae de mensura sortis*, translated into English in *Exposition of a new theory on the measurement of risk*, Econometrica, **22**, 23-36, 1954. In his 1812 book, Laplace formalizes the definition of probability which was currently used in the 18th century.

In order to illustrate the classical definition of probability, suppose you consider a dichotomous variable only assuming two values in an experiment. This is the case when tossing a coin. Now, if you toss the coin you have two possible outcomes: H (for head) and T (for tails). The probability $\mathrm{P}(H)$ of getting H is given by the number of favourable outcomes, 1 here, divided by the total number of outcomes, 2 here, so that:

$$\mathrm{P}(H) = \frac{\text{\# of favourable outcomes}}{\text{\# of possible outcomes}} = \frac{1}{2}.$$

The classical definition of probability remains a good guideline to solve probability problems and to get the correct result in many cases. The task of finding the probability of an event is reduced to a combinatorial problem. One must enumerate and count all the favourable cases as well as all the possible cases.

C. The classical definition in practice

In order to use the classical definition, one should be able to list favourable outcomes as well as the total number of possible outcomes of an experiment. Not always is this directly possible. Suppose you want to know which is the probability of exactly getting two heads in three tosses of a coin. There 8 possible cases: $(TTT, TTH, THT, HTT, HHT, HTH, THH, HHH)$ of which 3 contain 2 heads. Then the required probability is 3/8. If you consider 1o tosses of a coin, there are already 1024 possible cases and listing them all becomes boring. The fundamental counting principle comes into rescue.

> **Definition** (Fundamental counting principle) for a finite sequence of decisions, the number of ways to make these decisions is the product of the number of choices.

> **Example** In the case discussed above there are 3 decisions in a sequence (choosing H or T for three times) and there are 2 choices for every decisions (H or T). Thus, the total number of decisions is $2^3 = 8$.

Based on the fundamental counting principle, one gets the number of dispositions, permutations, combinations and combinations without repetition for N objects.

> **Example** (Dispositions with repetition) Suppose you want to choose an object n times out of N objects. The total number of possible choices is N each time and, based on the fundamental counting principle, one gets that there are N^n possible choices.

Example (Permutations) Now you want to pick an object out of N, remove it from the list of objects and go on until all the objects are selected. For the first decision you have N choices, for the second decisions $N-1$ and so on until the Nth decision where you just have 1 choice. As a consequence of the fundamental counting principle, the total number of possible decisions is $N!$.

Example (Dispositions without repetition) This time you are interested in selecting n objects out of N with $n \leq N$, but you are also interested in the order of the selected items. The first time you have N choices, the second time $N-1$ and so on, until the nth time where you have $N-n+1$ choices left. Then, the total number of possible decisions is $N(N-1)\cdots(N-n+1) = N!/n!$.

Example (Combinations) You have a list of N objects and you want to select n objects out of them with $n \leq N$, but you do not care about their order. Any ordered sublist with the desired n objects can be included in $(N-n)!$ lists with the remaining $N-n$ objects and there are $N!/n!$ of these sublists. Therefore, this time, the total number of possible decisions (possible way of selecting n objects out of N irrespective of their order) is $N!/(n!(N-n)!)$. This is a very useful formula and there is a special symbol for the so-called *binomial coefficient*:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

Indeed, these coefficients appear in the expansion of the Nth power of a binomial:

$$(p+q)^N = \sum_{n=0}^{N} \binom{N}{n} p^n q^{N-n},$$

where

$$\binom{N}{0} = \binom{N}{N} = 1$$

as a consequence of the definition $0! := 1$.

Example (Combinations with repetition) Suppose you are interested in finding the nuber of ways of allocating N objects into n boxes, irrespective of the names of the objects. Let the objects be represented by crosses, $\times$, and the boxes by the following symbol: $|\cdots|$. A particular configuration with the first box containing two objects and the last box empty is represented by $|\times\times|\cdots||$. As a further example, consider the case with two boxes and three objects with two objects in the first box and one object in the second box. This is $| \times \times | \times |$. Now, the total number of symbols is $N+n+1$ of which 2 are always fixed as the first and the last symbols must be a $|$. Of the remaining $N+n+1-2 = N+n-1$ symbols, N can be arbitrarily chosen to be crosses. The number of possible choices is then given by the binomial factor $\binom{N+n-1}{N}$.

Example (Tossing coins revisited) Let us consider once again the problem presented at the beginning of this subsection. This was: what is the probability of exactly finding two heads out of three tosses of a coin? Now, the problem can be generalized: what is the probability of exactly finding n heads out of N tosses of a coin ($n \leq N$)? The total number of possible outcomes is 2^N as there are 2 choices the first time, two the second and so on until two choices for the Nth toss. The number of favourable outcomes is given by the number of ways of selecting n places out of N and putting a head there and a tail elsewhere. Therefore

$$\mathrm{P}(\text{exactly } n \text{ heads}) = \binom{N}{n}\frac{1}{2^N}.$$

D. Circularity of the classical definition

Even if very useful for practical purposes, the classical definition suffers of circularity. In order to justify this statement, let us re-write the classical definition: *the probability of an event is given by the number of favourable outcomes divided by the total number of possible outcomes.* Now consider a particular outcome. In this case, there is only 1 favourable case, and if r denotes the total number of outcomes, one has

$$\mathrm{P}(\text{outcome}) = \frac{1}{r}.$$

This equation is the same for any outcome and this means that all the outcomes have the same probability. Therefore, in the classical definition, there seems to be no way of considering *elementary* outcomes with different probability and the equiprobability of all the outcomes is a sort of hidden assumption of the definition. A difficulty with equiprobability already arises in the case of dichotomous variables, where one of the outcomes could have a different probability with respect to the other outcome. The usual example is the unbalanced coin. If the hidden assumption is made explicit, then one immediately sees the circularity as probability is used to define itself: *the probability of an event is given by the number of favourable outcomes divided by the total number of possible outcomes assumed equiprobable.* In summary, if the equiprobability of outcomes is not mentioned, it becomes an immediate consequence of the definition and it becomes impossible to deal with non-equiprobable outcomes. If, on the contrary, the equiprobability is included in the definition as an assumption, then the definition becomes circular.

A possible way out from circularity was suggested by J. Bernoulli and adopted by Laplace; it is the so-called indifference principle. According to this principle, if one has no reason to assign different probabilities to a set of exhaustive and mutually exclusive events (called outcomes so far), then these events must be considered as equiprobable. For instance, in the case of the coin, in the absence of further indication, one has the following set of equations

$$\mathrm{P}(H) = \mathrm{P}(T)$$

and

$$\mathrm{P}(H) + \mathrm{P}(T) = 1$$

yielding $\mathrm{P}(H) = \mathrm{P}(T) = 1/2$, where the outcomes H and T are exhaustive (all the possible cases) and mutually exclusive (if one obtains H, one cannot have T at the same time).

The principle of indifference may seem a beautiful solution, but it leads to several problems and paradoxes identified by J. M. Keynes, by J. von Kries in *Die Prinzipien der Wahrscheinlichkeitsrechnung* published in 1886 and by Bertrand in his *Calcul des probabilités* of 1907. Every economist knows the *General theory*, but few are aware of *A treatise on probability*, a book published by Keynes in 1921 where one can find one of the first attempts to present probability axioms. Let us now see some of the paradoxes connected with the principle of indifference. Suppose that one does not know anything on a book. Therefore the probability of the statement *this book has a red cover* is the same as the probability of its negation *this book has not a red cover.* Again here one has a set of exhaustive and mutually exclusive events, whose probability is 1/2 according to the principle of indifference. However, as nothing is known on the book, the same considerations can be repeated for the statements *this book has a green cover*, *this book has a blue cover*, etc.. Thus, each of these events turns out to have probability 1/2, a paradoxical result. This paradox can be avoided if one further knows that the set of possible cover colours is finite and made up of, say, r

elements. Then the probability of *this book has a red cover* becomes $1/r$ and the probability of *this book has not a red cover* becomes $1 - 1/r$. Bertrand's paradoxes are subtler and they make use of the properties of real numbers. Already with integers, if the set of events is countable, the indifference principle leads to a distribution where every event has zero probability as $\lim_{r \to \infty} 1/r = 0$ but where the sum of these zero probabilities is 1, a puzzling result which can be dealt with using measure theory. The situation becomes worse if the set of events is infinite and non-countable. Following Bertrand, let us consider a circle and an equilateral triangle inscribed in the circle. What is the probability that a randomly selected chord is longer than the triangle side? Two possible answers are:

1. One of the extreme points of the chord can indifferently lie in any point of the circle. Let us then assume that it coincides with a vertex of the triangle, say vertex A. Now the chord direction can be selected by chance and the chord is longer than the side of the triangle only if its other extreme point lies on the circle arc opposite to vertex A. The triangle defines three circle arcs of equal length, this means that the required probability is 1/3.

2. Random selection of a chord is equivalent to random selection of its central point. In order for the chord to be longer than the triangle side, the distance of its central point from the centre of the circle must smaller than one-half of the circle radius. Then the area to which this point must belong is 1/4 of the circle area and the corresponding probability turns out to be 1/4 instead of 1/3.

There are other possible ways of avoiding or circumventing the circularity of the classical definition. One is the frequentist approach, where probabilities are identified with measured frequencies of outcomes in repeated experiments. Another solution is the subjectivist approach, particularly interesting for economists as probabilities are there defined in terms of *rational bets*.

E. Frequentism

The principle of indifference introduces a subjective element in the evaluation of probabilities. If, in the absence of any reason, one can assume equiprobable events, then if there are specific reasons one can make another assumption. Then probability assignments depend on one's state of knowledge on the investigated system. Empiricists opposed similar views and tried to focus on the outcomes of real experiments and to define probabilities in terms of frequencies. Roughly speaking this line of thought can be explained as follows with the example of coin tossing. According to frequentists the probability of H can be approximated by repeatedly tossing a coin, by recording the sequence of outcomes $HHTHTTHHTTTH\cdots$, counting the number of H and dividing for the total number of trials

$$\mathrm{P}(H) \sim \frac{\#\text{ of }H}{\#\text{ of trials}}.$$

The ratio on the right hand side of the equation is the *empirical frequency* of the outcome H, a useful quantity in descriptive statistics. Now, this ratio is never equal to 1/2 and the frequentist idea is to extrapolate the sequence of trials to infinity and to define the probability as

$$\mathrm{P}(H) = \lim_{\#\text{ of trials}\to\infty} \frac{\#\text{ of }H}{\#\text{ of trials}}.$$

This is the preferred definition of probability in several textbooks introducing probability and statistics to natural scientists and in particular to physicists. Probability becomes a sort of measurable quantity that does not depend on one's state of knowledge, it becomes *objective* or, at least, *intersubjective*. Moreover, Kolmogorov himself was a supporter of the frequentist point of view and his works on probability theory have been very influential. The naive version of frequentism presented above cannot be a solution to the problems discussed before. Indeed, the limit appearing in the definition of probability is not the usual limit defined in calculus for the convergence of a series. There is no formula for the number of heads out of N trials and nobody can toss a coin for an infinite number of times. Having said that, one can notice that similar difficulties are present when one wants to define real numbers as limits of Cauchy sequences of rational numbers following Dedekind. This solution to the objection presented above has been proposed by Von Mises, who starting from 1919 tried to develop a rigorous frequentist theory of probability. His first memoir was *Grundlagen der Wahrscheinlichkeitsrechnung* published in *Mathematischen Zeitschrift*. Subsequently, he published a book, *Wahrscheinlichkeit Statistik und Wahrheit. Einführung in die neue Wahrcheinlichkeitslehere und ihre Anwendungen* (Probability, statistics and truth. Introduction to the new probability theory and its applications). There are several difficulties in Von Mises' theory, but they can be solved and, in principle, a rigorous frequentist probability theory can be developed. Another method to circumvent the problems related to infinite trials has been explored by the late Kolmogorov, who developed a *finitary* frequentist theory of probability connected with his theory of information and computational complexity.

One of the reasons of the success of the frequentist approach was related to the social success among statisticians and natural scientists of the methods developed by K. Pearson and R.A. Fisher who were strong supporters of frequentism. These methods deeply influenced the birth of Econometrics, the only branch of Economics (except for Mathematical Finance) making extensive use of probability theory.

The main objection to frequentism is that most events are not repeatable and in this case, it is impossible, even in principle, to apply a frequentist definition of probability based on frequencies simply because these frequencies cannot be measured at all. Von Mises explicitly excluded these cases from his theory. In other words, given an event that is not repeatable such as *tomorrow it will rain*, it is a nonsense to ask for its probability. Notice that virtually all of Economics falls outside the realm of repeatability. If one were to fully accept this point of view, most applications of probability and statistics to Economics (including most of Econometrics) would become meaningless. Incidentally, the success of frequentism could explain why there are so few probabilistic models in theoretical Economics.

F. Subjectivism

Frequentism wants to eliminate the subjective element present in the indifference principle. On the contrary, subjectivism accepts this element and amplifies it by defining probability as the degree of belief that each individual assigns to an event. This event need not to occur in the future and it is not necessary that the event is repeatable. Being *subjective*, the evaluation of probability may differ from individual to individual. However, any individual must assign his/her probabilities in a coherent way, so that, for instnce, a set of exhaustive and mutually exclusive events has probabilities summing up to 1. Moreover, if two individuals share the same knowledge, their probability assignments must coincide. This point of view is particularly appealing for theoretical Economists as, in subjectivism, individuals are rational agents and their assignment of probabilities follow what is known as *normative theory* in the theory of choice. Indeed, probabilities are related to bets. The evaluation of probabilities in terms of bets was independently proposed by Frank Plumpton Ramsey and by the Italian statistician Bruno de Finetti. This is the only case this author knows where an Italian scientist who published his results after an Anglo-saxon counterpart is better known within the scientific community. Ramsey died when he was 26 years old in 1930, whereas De Finetti died in 1985 at the age of 89. Ramsey published his results in 1926 in his notes on *Truth and probability* and De Finetti wrote his essay on the logical foundations of probabilistic reasoning (*Fondamenti logici del ragionamento probabilistico*) in 1930. From 1930 to 1985, De Finetti had a lot of time to further develop and publicize his views and he also published many important papers on statistics and probability, including an important theorem on exchangeable sequences known as De Finetti's theorem in the probabilistic literature.

In particular, De Finetti presented an operational procedure to define probabilities. Suppose that if the event A takes place a rational individual will be given 1 Euro and 0 Euro if A does not take place. This rational individual is now required to evaluate the probability of A taking place. To this purpose, he/she has to choose an amount x such that he/she will lose $(1-x)^2$ Euros if A occurs and x^2 Euros if A does not occur. Now we can define a (random) variable L representing the loss and depending on whether A takes place. If A occurs one has $L = (1-x)^2$, if A does not occur then $L = x^2$. Now, if $p = \mathrm{P}(A)$ is the probability of occurrence for the event A, the expected value of the loss L is

$$\mathbb{E}[L] = p(1-x)^2 + (1-p)x^2,$$

so that

$$\mathbb{E}[L] = p - 2px + x^2.$$

The analysis of this quadratic function of x shows that the expected loss is minimal for $x = p = \mathrm{P}(A)$ and our rational agent must choose $x = p$ in order to minimize his/her loss. The extensions of this analysis to a finite set of mutually exclusive and exhaustive events is straightforward. For years, in Rome Bruno De Finetti organized a lottery based on these ideas to forecast Sunday football (soccer for US readers) match outcomes among his students and colleagues.

An objection to this line of thought can be based on the results of empirical economics and psychology. It turns out that human beings are not able to correcly evaluate probabilities even in simple cases. By the way, this also happens to many students and scholars of probability trying to solve elementary exercises. Probability is highly counterintuitive and even if one can conceive a rational agent who can base his/her choices in the presence of uncertainty on perfect probabilistic calculations, this is not a human being.

G. Logical empiricism

There is a third way to the solution of the problems posed by probability theory and it is related to a phylosophical movement known as *logical empiricism* or *logical positivism.* In agreement with subjectivists, for logical empiricists it is meaningful to discuss the probability of single events. However, like frequentists, logical empiricists believe in objective probabilities. The early Keynes was following this philosophical movement in trying to axiomatize probability as a measure of the relation between propositions. In his *Tractatus Logico-Philosophicus* also Ludwig Wittgenstein suggested a connection between logic and probability, although his connections with positivism is controversial. Later, Rudolf Carnap published several paper which were collected in the volume *Logical Foundations of Probability.* As with the other approaches, a detailed analysis of the ideas based on logical empiricism is outside the scope of these notes. Nonetheless, it is useful to explore some elementary aspects of logical empiricism, because they are very useful for the solutions of exercises.

1. Boolean lattices

Both propositional logic and the elementary algebra of sets share an algebraic structure known as Boolean lattice (or Boolean algebra).

Definition (Boolean lattice) A Boolean lattice is an algebraic structure $(\mathcal{B}, 0, 1,', +, \cdot)$ where $\mathcal{B}$ is a set, $0 \in \mathcal{B}$, $1 \in \mathcal{B}$, $'$ is a unary operation on $\mathcal{B}$ (that is a function $\mathcal{B} \to \mathcal{B}$), and $+, \cdot$ are binary operations on $\mathcal{B}$ (that is functions $\mathcal{B}^2 \to \mathcal{B}$) satisfying the following axioms ($a, b, c \in \mathcal{B}$):

1. Associative property 1
$$a + (b + c) = (a + b) + c;$$
2. Associative property 2
$$a \cdot (b \cdot c) = (a \cdot b) \cdot c;$$

3. Commutative property 1

$$a + b = b + a$$

4. Commutative property 2

$$a \cdot b = b \cdot a;$$

5. Distributive property 1

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c);$$

6. Distributive property 2

$$a + (b \cdot c) = (a + b) \cdot (a + c);$$

7. Identity 1

$$a + 0 = a;$$

8. Identity 2

$$a \cdot 1 = a;$$

9. Property of the complement 1

$$a + a' = 1;$$

10. Property of the complement 2

$$a \cdot a' = 0.$$

The presence of constants excludes the possibility that a Boolean lattice is empty. Now, given a set Ω and the set of its subsets $\mathcal{P}\Omega$, one can identify $\mathcal{B}$ with $\mathcal{P}\Omega$, 0 with $\emptyset$, 1 with Ω, $'$ with the complement ^{c}, $+$ with the union $\cup$ and $\cdot$ with the intersection $\cap$. Direct inspection shows that $(\mathcal{P}\Omega, \emptyset, \Omega, ^c, \cup, \cap)$ is a Boolean lattice. In order to find the connection with propositional logic, one can consider the classes of equivalent propositions. They are a Boolean lattice where 0 is the class of contradictions, 1 the class of tautologies, $'$ corresponds to the logical connective NOT, $+$ corresponds to OR, and $\cdot$ corresponds to AND.

In probability theory, one can ask questions of the following kind: What is the probability of finding a K *or* a J in the first draw from a deck of 52 cards? If two dice are thrown what is the probability of an odd number on the first dice *and* an even number on the second one? What is the probability of *not* getting a one when a dice is thrown? Virtually all (formal or informal) probability theories include the following rules for combining elementary probabilities:

1. (Probability of mutually exclusive events) If A and B are mutually exclusive events/propositions then
$$\mathrm{P}(A\ OR\ B) = \mathrm{P}(A) + \mathrm{P}(B);$$
2. (Probability of independent events) If A and B are independent events/propositions then
$$\mathrm{P}(A\ AND\ B) = \mathrm{P}(A)\mathrm{P}(B);$$
3. (Probability of complementary event) If A is an event, the probability that A does not take place is
$$\mathrm{P}(NOT\ A) = 1 - \mathrm{P}(A).$$

Example (Disjoint events) What is the probability of finding a K *or* a J in the first draw from a deck of 52 cards? The two sets of favourable event: 4 K's and 4 J's have no common element. Therefore there are 8 favourable cases out of 52, which means that the sought probability is $8/52 = 2/13 \simeq 0.15$. Using rule 1 (Probability of mutually exclusive events), one gets the same result: $4/52$ is the probability of getting a K and $4/52$ is the probability of getting a J, therefore the answer is $4/52 + 4/52 = 8/52$.

Example (Independent events) If two dice are thrown what is the probability of an odd number on the first dice and an even number on the second one? There are 3 odd numbers on the first die and 3 even numbers on the second die for a total of 9 favourable case as a consequence of the fundamental counting principle. Accordingly, the total number of cases is 36 and the answer is $9/36 = 1/4$. Now, the probability of an odd number on the first die is 1/2 and the probability of an odd number on the second die is 1/2. The result of one die does not depend on the result of the other die, and one gets for the answer $0.5 \cdot 0.5 = 0.25$, an application of rule 2 (Probability of independent events).

Example (Probability of negation) What is the probability of not getting a one when a dice is thrown? Here there are 5 favourable cases out of 6 and therefore the sought probability is $5/6 \simeq 0.83$. This is equivalent to subtracting 1/6 (the probability of getting 1) from 1 according to rule 3 (Probability of complementary event): $1 - 1/6 = 5/6$.

H. Where do we stand?

In the previous sections, a short outline of the foundational problems of probability theory has been presented. Indeed, each solution proposed to cope with the circularity of the classical definition (the principle of indifference, frequentism, subjectivism, logical empiricism) has its own shortcomings. To be true, we have not discussed the logical approach, but as you will see in the next section, even the most successful theory developed by Kolmogorov does not contain any rule or guidance to assess probabilities, whereas it only contains rules to appropriately combine them once they are known. This is a problem with logical empiricism as well.

Understanding the meaning of probability is a difficult problem and it is still unsolved after centuries of research. Unfortunately, in the last decades, such problems have been left either to philosophers or to a small number of senior scientists at the end of their scientific careers. Many philosophers dealing with these problems are very good, but often they are not active mathematicians and their results are not known outside their circles. One of the limits of publish-or-perish science is that active scientists are left with virtually no time for speculation on foundational problems. Younger scientists who wish to devote themselves to foundations often put their careers at a stake. This situation is more than unfortunate because probability theory has applications in many fields of science and advances in foundational studies might help in clarifying many problems present in other disciplines incuding Economics.

II. KOLMOGOROV'S AXIOMS

In this section, the axiomatic foundation of probability theory will be introduced. After reading and studying this section you should be able to

- prove simple but relevant theorems starting from the axioms;
- define disjoint events and independent events;
- define conditional probabilities;
- explain some counterintuitive aspects of probability theory.

A. The probability space

Before listing the axioms, it is useful to introduce three mathematical objects on which the axioms will be based: the probability space Ω, the σ-field $\mathcal{F}$ of subsets of Ω and the probability P. Kolmogorov based its axioms on measure theory, therefore he introduced probability using a *measure space*. Indeed, a *probability space* is an instance of measure space. First of all one needs the definition of σ-field:

> **Definition** (σ-field): Given a set Ω, a class $\mathcal{F}$ of subsets of Ω is a σ-field if:
>
> 1. the empty set $\emptyset$ is in $\mathcal{F}$;
> 2. if a set A is in $\mathcal{F}$ then also its complement A^c is in $\mathcal{F}$;
> 3. if $A_1, A_2, \ldots$ is a countable collection of sets belonging to $\mathcal{F}$ then also their countable union $\cup_i A_i$ and their countable intersection $\cap_i A_i$ are in $\mathcal{F}$.

In other words a σ-field is a collections of subsets of Ω closed with respects to the operations of complement, countable union and countable intersection. Moreover, it contains the empty set $\emptyset$ and also the set Ω as $\emptyset^c = \Omega$. Then, it is possible to define the probability space as follows:

> **Definition** (Probability space): A probability space is a triple $(\Omega, \mathcal{F}, \mathrm{P})$ where Ω is a set, $\mathcal{F}$ is a σ-field of subsets of Ω and P is a function from Ω to the real interval $[0, 1]$ such that:

1. $\mathrm{P}(A) \geq 0$ for any A in $\mathcal{F}$;
2. $\mathrm{P}(\Omega) = 1$;
3. if $A_1, A_2, \ldots$ is a countable collection of mutually disjoint sets belonging to $\mathcal{F}$ then $\mathrm{P}\left(\cup_i A_i\right) = \sum_i \mathrm{P}(A_i)$.

Remark(Countably additive, non-negative measure) P is an instance of countably additive and non-negative measure. A function μ from a σ-field to $\mathbb{R}$ is a countably additive, non-negative measure if:

1. $0 \leq \mu(A) \leq \infty$ for each set A belonging to the σ-field $\mathcal{F}$;
2. $\mu(\emptyset) = 0$;
3. if if $A_1, A_2, \ldots$ is a countable collection of mutually disjoint sets belonging to $\mathcal{F}$ then $\mu\left(\cup_i A_1\right) = \sum_i \mu(A_i)$.

B. Elementary consequences of the axioms

In order to show that P satisfies the axioms of countably additive, non-negative measure, it suffices to show that $\mathrm{P}(\emptyset) = 0$. This is an immediate corollary of the following theorem.

Theorem (Probability of complement) For each A in $\mathcal{F}$ one has

$$\mathrm{P}(A^c) = 1 - \mathrm{P}(A). \tag{1}$$

Proof. For each set A in $\mathcal{F}$ (indeed for each subset of Ω) one has that $A \cup A^c = \Omega$ and $A \cap A_c = \emptyset$, therefore as a consequence of the third axiom in (Probability space), the following chain of equalities holds true

$$1 = \mathrm{P}(\Omega) = \mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c)$$

hence $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$. □

Corollary (Probability of empty set) $\mathrm{P}(\emptyset) = 0$.

Proof. One has that $\Omega^c = \emptyset$ and $\mathrm{P}(\Omega) = 1$ according to the second axiom in (Probability space). Therefore, one has $\mathrm{P}(\emptyset) = \mathrm{P}(\Omega^c) = 1 - \mathrm{P}(\Omega) = 1 - 1 = 0$. □

Definition (Events) The sets in $\mathcal{F}$ are called events.

The axioms give a rule to compute the probability of the union of mutually disjoint events when the probability of each event is known. Finding the probability of the complement is an immediate consequence of the axioms. One still has to find rules to compute the probability of the union of non-disjoint events as well as the probability of the intersection of events. Indeed, the two problems are correlated. The following theorem shows how to compute the probability of the intersection of two events.

Theorem (Probability of union) For each couple of sets A and B in $\mathcal{F}$, one has

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B). \tag{2}$$

Proof. $A \cup B$ can be written as the union of three disjoint sets:

$$A \cup B = (A \cap B^c) \cup (A^c \cap B) \cup (A \cap B),$$

so that

$$\mathrm{P}(A \cup B) = \mathrm{P}(A \cap B^c) + \mathrm{P}(A^c \cap B) + \mathrm{P}(A \cap B). \tag{3}$$

Moreover, one has that A and B can be written as union of the following disjoint sets

$$A = (A \cap B^c) \cup (A \cap B)$$

and

$$B = (A^c \cap B) \cup (A \cap B)$$

which means that

$$\mathrm{P}(A \cap B^c) = \mathrm{P}(A) - \mathrm{P}(A \cap B)$$

and

$$\mathrm{P}(A^c \cap B) = \mathrm{P}(B) - \mathrm{P}(A \cap B).$$

Replacing these equations in (3) yields the thesis (2). □

This result can be extended to the union of an arbitrary number of events. However, it is more convenient to present a derivation of the general result after introducing random variables and indicator functions in the next section.

C. Conditional probabilities

As mentioned above, the probability of a generic union of two sets depends on the probability of the intersection. In Kolmogorov's theory, the probability of the intersection of two events is related to conditional probabilities via a definition:

Definition (Conditional probability) For each couple of sets A and B in $\mathcal{F}$, the conditional probability of A given B is defined as follows

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}, \tag{4}$$

whereas, in other theories, equation (4) is derived as a theorem from a different set of axioms. Bayes' rule is an immediate consequence of this definition.

Theorem (Bayes' rule) If $\mathrm{P}(A) \neq 0$ and $\mathrm{P}(B) \neq 0$, the conditional probabilities $\mathrm{P}(A|B)$ and $\mathrm{P}(B|A)$ are related as follows

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\mathrm{P}(B)}{\mathrm{P}(A)}. \tag{5}$$

Proof. From the definition of conditional probability in (4), one has

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

and

$$\mathrm{P}(A \cap B) = \mathrm{P}(B|A)\mathrm{P}(A),$$

hence

$$\mathrm{P}(B|A)\mathrm{P}(A) = \mathrm{P}(A|B)\mathrm{P}(B)$$

and the thesis follows. □

Given a partition of Ω into a family of mutually disjoint sets $\{E_i\}_{i=1}^n$, one can derive the so-called theorem of *total probability*:

Theorem (Total probability) If $\{E_i\}_{i=1}^n$ is a family of mutually disjoint events, $E_i \cap E_j = \emptyset$ for any $i \neq j$, such that $\Omega = \cup_{i=1}^n E_i$ then for any A in $\mathcal{F}$

$$\mathrm{P}(A) = \sum_{i=1}^{n} \mathrm{P}(A|E_i)\mathrm{P}(E_i). \tag{6}$$

Proof. The following chain of equalities holds true

$$A = A \cap \Omega = A \cap (\cup_{i=1}^{n} E_i) = \cup_{i=1}^{n} (A \cap E_i),$$

then A is written as the union of mutually disjoint sets and

$$\mathrm{P}(A) = \mathrm{P}\left(\cup_{i=1}^{n} (A \cap E_i)\right) = \sum_{i=1}^{n} \mathrm{P}(A \cap E_i) = \sum_{i=1}^{n} \mathrm{P}(A|E_i)\mathrm{P}(E_i)$$

yielding the thesis. □

Combining Bayes' rule and total probability, the following interesting result follows

Corollary (Bayes' theorem) If $\{E_i\}_{i=1}^{n}$ is a family of mutually disjoint events, $E_i \cap E_j = \emptyset$ for any $i \neq j$, such that $\Omega = \cup_{i=1}^{n} E_i$ then for any A in $\mathcal{F}$ and for any j

$$\mathrm{P}(E_j|A) = \frac{\mathrm{P}(A|E_j)\mathrm{P}(E_j)}{\sum_{i=1}^{n} \mathrm{P}(A|E_i)\mathrm{P}(E_i)}. \tag{7}$$

Proof. Bayes' rule states that

$$\mathrm{P}(E_j|A) = \frac{\mathrm{P}(A|E_j)\mathrm{P}(E_j)}{\mathrm{P}(A)}$$

total probability that

$$\mathrm{P}(A) = \sum_{i=1}^{n} \mathrm{P}(A|E_i)\mathrm{P}(E_i).$$

The thesis follows by direct substitution of the second equation into the first one. □

Bayes' theorem has a rather infamous interpretation, where $\mathrm{P}(E_j|A)$ is considered as the probability of the *cause* E_j given the *effect* A. However, in general, a causal interpretation of conditional probabilities is *wrong* and/or *misleading*. This error is rather common in Physics. Conditional probability is an elusive concept and conditioning reflects a *logical* or *informational* relationship between events, not always a causal relationship. The following classical examples illustrate this point.

Example (Retrodiction) Suppose you have a deck of 52 cards. Two cards are drawn without replacement and in a sequence. If the first card is a K, what is the probability that the second card is a K? Almost everybody acquainted with elementary probabilities gives the correct answer 3/51. Now, imagine you know that the second card is a K without having any piece of information on the first card, then what is the probability of drawing a K at the first draw? The correct answer is again 3/51. Why?

Example (Monty Hall quiz) You are in front of three closed doors, numbered from 1 to 3. Two of them are empty and the third one contains a valuable prize, let say a luxury car. You do not know where the car is. The quiz master (who perfectly knows where the prize is) asks you to choose one door. You choose door number 1. Now the quiz master opens door 3 (it is empty) and gives you the possibility of changing your choice. What is better for you? Changing your mind and selecting door number 2 or keeping your first decision? It turns out that the probability of winning when changing is 2/3 whereas if you do not change your choice, you win only 1 time out of 3. Why? This game has become very important and popular in experimental psychology and experimental economics. It is very difficult to give the correct answer at the first attempt. One reason could be that it is difficult to understand all the details correctly. P.R. Mueser and D. Graberg give a somewhat redundant description of the game. Their paper, *The Monty Hall Dilemma Revisited: Understanding the Interaction of Problem Definition and Decision Making* (University of Missouri Working Paper 99-06), is on-line: `http://econwpa.wustl.edu:80/eps/exp/papers/9906/9906001.html`. Try to read it and see if you can figure out a way of justifying the solution.

Luckily, as readers of these notes, you may never have been exposed to textbooks on theoretical physics dealing with quantum mechanics as well as to textbooks on introductory experimental physics. Therefore, you have better chances of understanding the informational and subjective meaning of conditional probabilities without having to care of wavefunction collapses and similar nonsense.

Remark (Numerical values of probabilities) Notice that in Kolmogorov's axiomatic theory, there is no algorithm, no rule helping in determine the actual value of probabilities. For this reason, for real world problems one has some freedom in order to assign probabilities. One usually uses a *heuristic mixture* of the classical definition and of the subjective definition. The mildly measure theoretical notions of the next sections, where random variables and some stochastic processes are introduced, are very useful as it turns out that all the apparatus of calculus can be used for studying probabilistic problems.

Remark (Probabilities on a countable space) Let us now assume that the space Ω is countable and that the σ-field $\mathcal{F}$ coincides with the class of all subsets of Ω: $\mathcal{F} = \mathcal{P}\Omega$.

Definition (Atoms) The elements $\{\omega_i\}_{i=1}^{\infty}$ of Ω are called atoms.

Theorem A probability on Ω is characterized by its values $p_i = \mathrm{P}(\omega_i)$ on the atoms $\omega_i \in \Omega$.

Proof. As we have seen for the total probability theorem, for any event A, one has that $A = A \cap \Omega = A \cap (\cup_{i=1}^{\infty} \omega_i) = \cup_{\omega_i \in A} \omega_i$, therefore $P(A) = \sum_{\omega_i \in A} \mathrm{P}(\omega_i) = \sum_{\omega_i \in A} p_i$. □

Theorem Let Ω be a countable set and $\{\omega_i\}_{i=1}^{\infty}$ one of the possible enumerations of the atoms of Ω. If $\{p_i\}_{i=1}^{\infty}$ is a sequence of real numbers, then there exists a unique probability measure P such that $\mathrm{P}(\omega_i) = p_i$ if and only if $p_i \geq 0$ and $\sum_{i=1}^{\infty} p_i = 1$.

Proof. If $0 \leq \mathrm{P}(\omega_i) = p_i$ then one has $1 = \mathrm{P}(\Omega) = \mathrm{P}(\cup_{i=1}^{\infty} \omega_i) = \sum_{i=1}^{\infty} \mathrm{P}(\omega_i) = \sum_{i=1}^{n} p_i$. Conversely, if the p_i satisfy $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$, one can define for any event A the measure $\mathrm{P}(A) = \sum_{\omega_i \in A} p_i$. This measure satisfies the axioms of probability. As for countable additivity, one can observe that $\sum p_i$ is a positive absolutely convergent series and it is possible to add disjoint partial sums to get the same result as the total sum. □

III. RANDOM VARIABLES

In this section, random variables will be introduced as measurable functions from a probablity space to the set of real numbers equipped with the structure of positive measurable space $(\mathbb{R}, \mathcal{B}, \mu)$ where $\mathcal{B}$ is the Borel σ-field, that is the smallest σ-field containing open subsets of $\mathbb{R}$ and μ is a positive measure on $\mathbb{R}$ such that $\mu(\mathbb{R}) = 1$.

After reading this section, you should be able to:

- properly define random variables;
- define the probability distribution and probability density for a random variable;
- define the expectation of a random variable;
- derive elementary properties of the expectation;
- define the variance of a random variable;
- define the moments of a random variable;
- understand the principles of multivariate analysis;
- define the conditional expectation of a random variable;
- derive the properties of conditional expectation;
- understand the concept of sequences of random variables and the different convergence definitions;
- understand some important convergence theorems such as the weak and strong laws of large numbers as well as the central limit theorem.

A. More on σ-fields

In order to define random variables, one first need to define the so-called σ-field generated by a class of subsets $\mathcal{E}$ of a set $\mathcal{X}$. Roughly speaking, this is necessary due to some problems in measure theory where there is no measure satisfying the proper axioms and for which all the sets in the class of all subsets of $\mathcal{X}$, denoted by $\mathcal{P}\mathcal{X}$, are measurable. Therefore one has to build measurable classes of subsets of $\mathcal{X}$.

> **Definition** (σ-field generated by $\mathcal{E}$) Given a set $\mathcal{X}$ and a class of its subsets $\mathcal{E}$, the σ-field generated by $\mathcal{E}$, $\sigma(\mathcal{E})$, is the smallest σ-field on $\mathcal{X}$ containing all sets of $\mathcal{E}$.

To prove that this definition is meaningful, it is necessary to show that if $\{\mathcal{F}_i\}_{i\in\mathcal{I}}$ is a family of σ-fields on $\mathcal{X}$, then their intersection $\cap_{i\in\mathcal{I}}\mathcal{F}_i$ (the collection of all sets belonging to every $\mathcal{F}_i$) is also a σ-field on $\mathcal{X}$. Moreover, one should also notice that for each $\mathcal{E}$ there is at least one σ-field containing all the sets in $\mathcal{E}$ and this is indeed $\mathcal{P}\mathcal{X}$. As for the first requirement, one can notice that $\emptyset$ is contained in every $\mathcal{F}_i$ and then it is also contained in $\cap_{i\in\mathcal{I}}\mathcal{F}_i$. Then if $A \in \cap_{i\in\mathcal{I}}\mathcal{F}_i$ also its complement A^c must be in the intersection as it is contained in every $\mathcal{F}_i$. The same applies to a countable union or intersection of sets in $\cap_{i\in\mathcal{I}}\mathcal{F}_i$. Therefore the definition is meaningful.

Example (σ-field generated by a finite set) Consider the case where $\mathcal{X}$ is a finite set, say a finite set containing 5 objects: $\mathcal{X} = \{a, b, c, d, e\}$. Now suppose that the class $\mathcal{E}$ is made up of two sets $E_1 = \{a, b, c\}$ and $E_2 = \{b, d, e\}$. In this case, the σ-field generated by $\mathcal{E}$ can be generated by a direct application of the axioms. It must include $\emptyset$ and $\mathcal{X}$, $F_1 = E_1^c = \{d, e\}$ and $F_2 = E_2^c = \{a, c\}$ as well as all the possible unions and intersections of these sets and their complements, $F_3 = E_1 \cap E_2 = \{b\}$ and $F_4 = F_1 \cup F_2 = E_1^c \cup E_2^c = \{a, c, d, e\}$. Further unions and intersections create no new sets. In summary, one has

$$\sigma(E_1, E_2) = \emptyset, E_1 = F_2 \cup F_3, E_2 = F_1 \cup F_3, F_1, F_2, F_3, F_4 = F_1 \cup F_2, \mathcal{X}.$$

F_1, F_2 and F_3 are mutually disjoint sets whose union is $\mathcal{X}$ and all the other sets in $\sigma(E_1, E_2)$ are union or intersections of these sets. They are called *atoms* of the σ-field.

Definition (Borel σ-field) The Borel σ-field, $\mathcal{B}(\mathbb{R})$, on $\mathbb{R}$ is the smallest σ-field generated by the open subsets of $\mathbb{R}$.

The Borel σ-field is generated by an infinite class of subsets of $\mathbb{R}$ and the direct procedure presented above cannot be used to characterize it. The so-called *generating class argument* can be used to prove that the Borel σ-field coincides with the σ-field generated by all the intervals of the kind $(-\infty, t]$ where $t \in \mathbb{R}$.

Theorem (Generating class argument) Let $\sigma(\mathcal{E})$ denote the σ-field generated by the intervals $\mathcal{E} = (-\infty, t]$ where $t \in R$, then $\sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$.

Proof. Let us denote by $\mathcal{G}$ the class of open subsets of the real line. First one proves that $\mathcal{E} \subseteq \sigma(\mathcal{G}) = \mathcal{B}(\mathbb{R})$, then, this means that $\sigma(\mathcal{G})$ is one of the σ-fields entering the definition of $\sigma(\mathcal{E})$ and one gets $\sigma(\mathcal{E}) \subseteq \mathcal{B}(\mathbb{R})$. Indeed, each interval $(-\infty, t]$ can be written as a countable intersection of open sets $(-\infty, t] = \cap_{n=1}^{\infty}(-\infty, t + n^{-1})$, then, the intervals $(-\infty, t]$ belong to the Borel σ-algebra, then $\mathcal{E} \subseteq \mathcal{B}(\mathbb{R})$ and, as a consequence $\sigma(\mathcal{E}) \subseteq \mathcal{B}(\mathbb{R})$. To prove that $\mathcal{B}(\mathbb{R}) \subseteq \sigma(\mathcal{E})$, one needs a representation of an open set as union, intersection or complement of sets in $\sigma(\mathcal{E})$. As any open set on the real line can be written as a countable union of open intervals, it suffices to discuss open intervals of the kind (a, b). They can be written as $(a, b) = (-\infty, b) \cap (-\infty, a]^c$ and one further has that $(-\infty, b) = \cup_{n=1}^{\infty}(-\infty, b - n^{-1}]$. This shows that $\mathcal{G} \subseteq \sigma(\mathcal{E})$ and, thus, $\mathcal{B}(\mathbb{R}) \subseteq \sigma(\mathcal{E})$. □

Remark (Non-uniqueness of the characterization) Notice that the characterization of the Borel σ-field is not unique.

B. Random variables

Based on the above discussion, it is possible to define measurable functions.

Definition ($\mathcal{A}\backslash\mathcal{B}$-measurable function) Let $\mathcal{X}$ be a set with its σ-field $\mathcal{A}$ and $\mathcal{Y}$ be another set with its σ-field $\mathcal{B}$, then a function or map $T : \mathcal{X} \to \mathcal{Y}$ is measurable if for each set $B \in \mathcal{B}$ the inverse image $A = \{x \in \mathcal{X} : \ T(x) \in B\}$ belongs to $\mathcal{A}$.

Remark (Notation for the inverse image) Notice that the inverse image of B with respect to the function T is often denoted by $T^{-1}(B)$ or even $T^{-1}B$. One should be careful not to mix up this notation with the reciprocal of a function!

And a specialization of the above definition leads to the definition of random variable.

Definition (Random variable) A random variable, X, is a $\mathcal{F}\backslash\mathcal{B}(\mathbb{R})$-measurable function where $\mathcal{F}$ is the σ-field of the probability space Ω and $\mathcal{B}(\mathbb{R})$ is the Borel σ-field.

After defining a random variable as a measurable function from Ω to $\mathbb{R}$, one needs a method to use integrals defined on $\mathbb{R}$ in order to compute probabilities. This is provided by the concept of *image measure*.

Definition (Image measure) Let μ be a measure on a σ-field $\mathcal{A}$ of subsets of $\mathcal{X}$ and let $T : \mathcal{X} \to \mathcal{Y}$ be an $\mathcal{A} \backslash \mathcal{B}$-measurable function where $\mathcal{B}$ is a σ-field of subsets of $\mathcal{Y}$. Then

$$\nu(B) := \mu(T^{-1}B) \tag{8}$$

defines a measure on $\mathcal{B}$ called the image measure of μ under T.

Remark (Image measure) To convince oneself that the image measure is a measure, one can use the following facts: $T^{-1}(B^c) = (T^{-1}B)^c$ and $T^{-1}(\cup_i B_i) = \cup_i T^{-1} B_i$.

Definition (Distribution) The distribution, P_X of a random variable X is the inverse image of the probability measure P. In other words

$$P_X(B) := \mathrm{P}(X^{-1}B), \tag{9}$$

where $B \in \mathcal{B}(R)$.

Remark (Distribution) The distribution of a random variable X, denoted by P_X, is a probability measure on the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$. in other words, it satisfies all the axioms of Kolmogorov's theory for Ω replaced by $\mathbb{R}$ and $\mathcal{F}$ replaced by $\mathcal{B}(\mathbb{R})$. This can be proved following the method outlined above for proving that the inverse image of a measure is also a measure.

The distribution of a random variable is a measure on $\mathbb{R}$; it is important to distinguish it from the so-called *distribution function* a.k.a. *cumulative distribution function*. As intervals of the kind $(-\infty, a]$ where $a \in \mathbb{R}$ belong to $\mathcal{B}(\mathbb{R})$, one can introduce the following definition:

Definition (Cumulative distribution function) The function

$$F_X(x) = P_X\{(-\infty, x]\} = P_X(X \leq x) \tag{10}$$

is called the cumulative distribution function of the random variable X.

The cumulative distribution function has the following properties

Theorem (Properties of the cumulative distribution function) Let X be a random variable and $F_X(x)$ its cumulative distribution function, then

1. $F_X(x)$ is increasing with $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$;
2. $F_X(x)$ is right-continuous.

Proof. 1. is a consequence of the following facts. P_X is a non negative measure and if $A \subseteq B$ then $P_X(A) \leq P_X(B)$; one can apply Dominated Convergence to the sequences $(-\infty, -n] \downarrow \emptyset$ and $(-\infty, n] \uparrow \mathbb{R}$ when $n \to \infty$. 2. also follows from Dominated Convergence applied to the sequence $(-\infty, x + n^{-1}] \downarrow (-\infty, x]$.

□

The complementary cumulative distribution function is an important concept in many applications. It is also known as the *survival function*, especially for positive random variables.

Definition (Complementary cumulative distribution function) The function

$$\Psi_X(x) = 1 - F_X(x) = 1 - P(X \leq x) = P(X > x) \tag{11}$$

is called complementary cumulative distribution function.

David Pollard writes in his book *A User's Guide to Measure Theoretic Probability* that the cumulative distribution function does not play any important role in modern probability theory, except for the study of order statistics and for a method for building measures on $\mathcal{B}(\mathbb{R})$ as images of Lebesgue measure. If one is interested in applications, it is hard to agree with Pollard. However, it is useful to present the so-called *quantile transformation* as it is the basis of many methods to generate pseudo-random numbers. Later in this section, also order statistics will be discussed.

Example (Quantile transformation) Suppose that $F(x)$ is a right-continuous increasing function on $\mathbb{R}$ such that $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$. Then, there is a probability measure P for which $P\{(-\infty, x]\} = F(x)$. It is possible to explicitly construct P by defining the *quantile function* $q(t) = \inf\{x : F(x) \geq t\}$ for $t \in [0,1]$. The set $\{x \in \mathbb{R} : \; F(x) \geq t\}$ is of the form $[\alpha, +\infty)$ because of right continuity of $F(x)$ and $\alpha = q(t)$. In general, one has that

$$F(x) \geq t \text{ if and only if } x \geq q(t) = \alpha.$$

Notice that equations such as $F(q(t)) = t$ are true only if $F(x)$ is continuous and strictly increasing. If m denotes the restriction of Lebesgue measure to the Borel σ-field in $(0,1)$ one has that

$$P\{(-\infty, x]\} = m\{t : \; q(t) \leq x\} = m\{t : \; t \leq F(x)\} = F(x);$$

the first equation is the definition of inverse measure, the second equality is a consequence of the identity of sets. Then if a random variable ξ is uniformly distributed in (0,1), its transform $q(\xi)$ has cumulative distribution function $F(x)$.

Many applied scientists are fond of another function: the *probability density function.* In order to define it, one has to assume continuity of $F_X(x)$. Physicists often mix up the probability density function with the probability distribution function.

Definition (Probability density function) If $F_X(x)$ is a continuous function, then the probability density function is its first derivative with respect to x:

$$p_X(x) = \frac{dF_X(x)}{dx}. \tag{12}$$

The following properties of the probability density function immediately follow from its definition and from theorems on integrals. They are given without proof.

Theorem (Properties of the probability density function)

The probability density function has the following properties

1. From the probability density function to the cumulative distribution function
$$F_X(x) = \int_{-\infty}^{x} p_X(u)\, du; \tag{13}$$
2. Normalization of the probability density function
$$\int_{-\infty}^{+\infty} p_X(u)\, du = 1; \tag{14}$$

3. If $B \in \mathcal{B}(\mathbb{R})$ then
$$P_X(B) = \int_B p_X(u)\,du; \tag{15}$$
4. Relationship between the probability density function and the complementary cumulative distribution function 1
$$p_X(x) = -\frac{d\Psi_X(x)}{dx}; \tag{16}$$
5. Relationship between the probability density function and the complementary cumulative distribution function 2
$$\Psi_X(x) = 1 - \int_{-\infty}^{x} p_X(u)\,du = \int_{x}^{+\infty} p_X(u)\,du. \tag{17}$$

Using Dirac's δ function, it is possible to use probability densities also when $F_X(x)$ is not continuous.

> **Remark** (Dirac's δ function) Let us consider a random variable whose cumulative distribution function is $F(x) = \theta(x)$ where $\theta(x) = 0$ for $x < 0$ and $\theta(x) = 1$ for $x \geq 1$ (incidentally, this function is called Heaviside θ-function or step function in Physics and Engineering). This is a legitimate cumulative density function as it is increasing, right-continuous, $\lim_{x\to-\infty}\theta(x) = 0$ and $\lim_{x\to+\infty}\theta(x) = 1$. However, it is not continuous everywhere. Its derivative vanishes for all $x \in \mathbb{R}$ except for $x = 0$ where the derivative does not exist. It is however possible to rigorously define a mathematical object (a functional) as the derivative of $\theta(x)$:
> $$\delta(x) := \frac{d\theta(x)}{dx};$$
> this object is called Dirac's delta function. It was formally introduced by physicist P.A.M. Dirac who used it without knowing its rigorous definition. The functional $\delta(x)$ is a map from the class of continuous functions on $\mathbb{R}$ to real numbers and maps every function $f(x)$ to its value in 0. One usually uses the following notation:
> $$\int_{-\infty}^{+\infty} f(x)\delta(x)\,dx = f(0).$$
> In particular, if $f(x) = 1$ on $\mathbb{R}$ one has that
> $$\int_{-\infty}^{+\infty} \delta(x)\,dx = 1;$$

an immediate important consequence is that for $c \in \mathbb{R}$

$$\int_{-\infty}^{+\infty} c\delta(x)\, dx = c;$$

moreover one has that for $a \in \mathbb{R}$

$$\int_{-\infty}^{+\infty} \delta(x-a)\, dx = 1.$$

In summary, $\delta(x)$ represent the probability density of a random variable assuming the value $x = a$ with probability 1 whereas $\delta(x-a)$ is the probability density of a random variable assuming the value $x = a$ with probability 1. Consider now a random variable X that assumes the value -1 with probability p and the value $+1$ with probability $q = 1-p$, based on the previous discussion, one can represent its probability density function as $p_X(x) = p\delta(x+1) + (1-p)\delta(x-1)$.

One often has to work with more than one random variable. The joint distribution and the joint cumulative distribution function and probability density function are the objects generalizing the previous concepts to random vectors.

Definition (Joint distribution) Let X and Y be two random variables defined on the same probability space. Then $T(\omega) = (X(\omega), Y(\omega))$ is a *random vector* and it is a (measurable) map from Ω to $\mathbb{R}^2$. The image measure T(P) on $\mathcal{B}(\mathbb{R}^2)$ is called the joint distribution of X and Y and it is often denoted by $P_{X,Y}$.

Definition (Cumulative joint distribution function) The cumulative joint distribution function $F_{X,Y}(x,y)$ is defined as $F_{X,Y}(x,y) = P_{X,Y}(X \leq x \cap Y \leq y)$. This relationship is often written as $F_{X,Y}(x,y) = P_{X,Y}(X \leq x, Y \leq y)$.

Definition (Joint distribution function) The joint distribution function $p_{X,Y}(x,y)$ is such that

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p_{X,Y}(u,w)\, dw\, du.$$

Definition (Marginals) The marginal distribution function $F_X(x)$ is defined as

$$F_X(x) = \lim_{y \to +\infty} F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{+\infty} p_{X,Y}(u,w)\, dw\, du;$$

the marginal distribution function $F_Y(y)$ is defined as

$$F_Y(y) = \lim_{x \to +\infty} F_{X,Y}(x,y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{y} p_{X,Y}(u,w)\, dw\, du;$$

the marginal probability density function $p_X(x)$ is defined as

$$p_X(x) = \int_{-\infty}^{+\infty} p_{X,Y}(x,w)\, dw;$$

the marginal probability density function $p_Y(y)$ is defined as

$$p_Y(y) = \int_{-\infty}^{+\infty} p_{X,Y}(u,y)\, du.$$

The above definitions can be straightforwardly extended to include any number of random variables. Based on the previous discussion, one can immediately derive joint distribution functions for independent random variables.

Definition (Independent random variables) Two random variables are independent if and only if their joint distribution function is the product of the marginals

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

as well as their joint probability density is the product of the marginals

$$p_{X,Y}(x,y) = p_X(x) p_Y(y)$$

It turns out that, in principle, the knowledge of the marginals is sufficient to characterize the joint distribution function.

Remark (Copulas) Let us consider two random variables X and Y and their joint cumulative distribution function, $F_{X,Y}(x,y)$. A theorem due to Sklar shows that there exists a function $C:\ [0,1]^2 \to [0,1]$ called the copula, such that

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

where $F_X(x)$ and $F_Y(y)$ are the two marginals. Moreover if the marginal cumulative distribution functions are continuous then the copula is unique. This result can be extended to an anrbitrary number of random variables. Notice that the independence copula is $C(u,v) = uv$. This result is important as, empirically, marginal cumulative distribution functions are easier to evaluate than joint probability distributions. Unfortunately, Sklar's result is not constructive and, for this reason, there are intense ongoing efforts to study the properties as well as the applications of several copulas.

Example (Order statistics) Consider a random variable X characterized by a probability density $p_X(x)$. Suppose to draw N real numbers out of the distribution P_X and to order them. What is the probability distribution of the k-th number x_k? The ordering procedure defines a new random variable X_k, the real number in the k-th position in the list $x_1, \ldots, x_k, \ldots, x_N$ with $x_1 \leq \ldots \leq x_k \leq \ldots \leq x_N$. Then

$$p_{X_k}(x) = N \binom{N-1}{k-1} p_X(x)[F_X(x)]^{k-1}[1 - F_X(x)]^{N-k}.$$

This result can be justified as follows. Consider a particular draw, the k-th number can be the first or the second, ..., or the N-th drawn and this accounts for the factor N in front of the density. $p_X(x)$ gives the probability of getting the value x, the binomial factor takes into account that one has to choose $k-1$ numbers out of $N-1$ that are smaller than x and $N-k+1$ numbers that are larger than x. Finally, assuming independent draws, $[F_X(x)]^{k-1} = [P_X(X \leq x)]^{k-1}$ is the probability that $k-1$ numbers are smaller or equal than x and $[1 - F(x)]^{N-k} = [P_X(X > x)]^{N-k}$ is the probability that $N-k$ numbers are larger than x.

C. Expectation of random variables

It is now time to define the expectation of a random variable, sometimes called the expected values as well as average or mean. It is, however, better to use the terms average or mean for the statistical estimate of the expectation.

Definition (Expectation) The expectation or expected value of a random variable X characterized by the distribution P_X is the Lebesgue integral

$$\mathbb{E}[X] = \int_{\mathbb{R}} x P_X(\mathrm{d}x); \tag{18}$$

It can be proved that one has

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \, \mathrm{d}F_X(x), \tag{19}$$

and if the probability density $p_X(x)$ exists at least in the generalized sense described above based on Dirac's δ-functions, one has

$$\mathbb{E}[X] = \int_{\mathbb{R}} x p_X(x) \, \mathrm{d}x = \int_{-\infty}^{+\infty} x p_X(x) \, \mathrm{d}x. \tag{20}$$

Here, as in all introductory probability texts, in practice, only definition (20) will be used. Expectation is a number characterizing the "typical" value of a random variable. The following properties, whose proof is left to the reader, are an immediate consequence of its definition.

Theorem (Elementary properties of the expectation)

1. (Multiplication by a scalar) If X is a random variable and $a \in \mathbb{R}$, then

$$\mathbb{E}[aX] = a\mathbb{E}[X];$$

2. (Sum of random variables) If $X_1, X_2, \ldots, X_N$ are N random variables on the same probability space, then

$$\mathbb{E}[X_1 + X_2 + \ldots + X_N] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \ldots + \mathbb{E}[X_N].$$

The expectation of a function of a random variable X can be written in terms of the distribution of X. The result below seems trivial, on the contrary it is an important theorem not so easy to prove.

Theorem (Fundamental expectation theorem) If $h(x)$ is a Borel measurable function on $\mathbb{R}$ and X is a random variable then

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x) P_X(\mathrm{d}x); \tag{21}$$

if $p_X(x)$ exists one can write

$$\mathbb{E}[h(X)] = \int_{-\infty}^{+\infty} h(x) p_X(x) \, \mathrm{d}x. \tag{22}$$

Remark (Trasformation of random variables) Given a random variable X, via a Borel measurable function $h(x)$, one can define the new random variable $Y = h(X)$ (a measurable function is needed to insure that Y is also $\mathcal{F}\backslash\mathcal{B}(\mathbb{R})$-measurable). Then one has $P_Y(B) = \mathrm{P}(Y^{-1}(B))$ where B is a Borel set. $C = h^{-1}(B)$ is still a Borel set given that $h(x)$ is Borel measurable. Therefore, the inverse function $Y^{-1}(B)$ is $Y^{-1}(B) = X^{-1}(h^{-1}(B)) = X^{-1}(C)$. In other words one has

$$P_Y(B) = P_X(C) = P_X(h^{-1}(B))$$

a relationship that can be used to build P_Y if P_X is known. In particular, if X has a continuous probability density, $p_X(x)$, and $Y = h(X)$ where h is a continuous function then

$$p_Y(y) = \sum_{x \in h^{-1}(y)} \frac{p_X(x)}{|h'(x)|},$$

where $h^{-1}(y)$ is the counter-image of y and $h'(x)$ the first derivative of $h(x)$. For a bijection

$$p_Y(y) = \frac{p_X(x)}{|h'(x)|}.$$

Similar inversion formulae exist for functions of several random variables.

Remark (Probability density for the sum of two independent random variables) An important transformation concerns the sum of two independent random variables X and Y defined on the same probability space. Let us assume that their probability density functions, $p_X(x)$ and $p_Y(y)$, exist at least in the generalized form and let us define $Z = X + Y$. One then has

$$p_Z(z) = \int_{u\in\mathbb{R}} \int_{w=z-u} p_{X,Y}(u,w)\,\mathrm{d}u\mathrm{d}w =$$
$$\int_{u\in\mathbb{R}} \int_{w\in\mathbb{R}} \delta(w-z+u) p_X(u) p_Y(w)\,\mathrm{d}u\mathrm{d}w = \int_{u\in\mathbb{R}} p_X(u) p_Y(z-u)\,\mathrm{d}u; \quad (23)$$

as there is no reason to privilege integration over X, one can prove that

$$p_Z(z) = \int_{u\in\mathbb{R}} p_X(u) p_Y(z-u)\,\mathrm{d}u = \int_{w\in\mathbb{R}} p_X(z-w) p_Y(w)\,\mathrm{d}w; \quad (24)$$

in other words, the probability density of the sum of two independent random variable is the *convolution* of the probability densities of the two random variables. This result can be extended to the sum of many independent random variables by repeated application of the rule for two random variables. If $X_1, \dots, X_N$ are N independent random variables with generalized probability density, then if $Z = \sum_{i=1}^N X_i$, one gets

$$p_Z(z) = \int_{u_N \in \mathbb{R}} \cdots \int_{u_1 \in \mathbb{R}} p_{X_N}(u_N) p_{X_{N-1}}(u_{N-1} - u_N) \dots p_{X_1}(z - u_1) \, \mathrm{d}u_N \dots \mathrm{d}u_1. \quad (25)$$

Now an important method will be introduced based on the so-called indicator function and promoted by Bruno De Finetti.

Definition (Indicator function) Let A be a set $A \subseteq \mathcal{X}$ and x an element of $\mathcal{X}$. Then the indicator function $\mathbb{I}_A(x)$ is

$$\mathbb{I}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases} \quad (26)$$

By means of the indicator function one transforms a Boolean algebra into a Boolean ring.

Theorem (Properties of indicator functions) Let A and B be two sets, then one has

1. (Indicator function of intersection)

$$\mathbb{I}_{A \cap B}(x) = \mathbb{I}_A(x) \mathbb{I}_B(x);$$

2. (Indicator function of union)

$$I_{A \cup B}(x) = \mathbb{I}_A(x) + \mathbb{I}_B(x) - \mathbb{I}_A(x) \mathbb{I}_B(x).$$

Proof. The two results can be proved by showing that the first and the right-hand side of the equations always coincide. As for the intersection, $\mathbb{I}_{A \cap B}(x)$ is 1 if and only if $x \in A$ and $x \in B$. In this case both $\mathbb{I}_A(x)$ and $\mathbb{I}_B(x)$ are 1 and their product is 1. In all the other possible cases, $\mathbb{I}_A(x)\mathbb{I}_B(x)$ vanishes. For what concerns the union, its indicator function is one if and only if $x \in A$ or $x \in B$. Now, if $x \in A$ and $x \notin B$ as well as if $x \notin A$ and $x \in B$, the sum $\mathbb{I}_A(x) + \mathbb{I}_B(x) - \mathbb{I}_A(x)\mathbb{I}_B(x)$ is 1 as either $\mathbb{I}_A(x)$ or $\mathbb{I}_B(x)$ are 1 and the other terms vanish. If $x \in A$ and $x \in B$, then again the right hand side of the equality is 1 as $\mathbb{I}_A(x) = 1$ and $\mathbb{I}_B(x) = 1$. In the remaining case, $x \notin A$ and $x \notin B$, all the three terms vanish. □

Remark (Indicator function and random variables) Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and let $A \subseteq \mathcal{F}$ be an event; the correspondence between A and its indicator function $\mathbb{I}_A$ establishes a correspondence between the σ-field $\mathcal{F}$ and a subset of the collection of random variables on Ω. The expectation maps random variables into real numbers in such a way that $\mathbb{E}[\mathbb{I}_A] = \mathrm{P}(A)$. This can be directly seen if one uses $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$, the Borel probability space where X is a random variable and P_X is the image measure. Let now $B \subseteq \mathcal{B}(\mathbb{R})$ be a Borel set and let $\mathbb{I}_B(x)$ be its indicator function. Then, assuming for simplicity the existence of a (possibly generalized) probability density function $p_X(x)$, one has the following chain of equalities

$$\mathbb{E}[\mathbb{I}_B(x)] = \int_{\mathbb{R}} \mathbb{I}_B(u) p_X(u)\,\mathrm{d}u = \int_B p_X(u)\,\mathrm{d}u = P_X(B) = \mathrm{P}(X^{-1}B).$$

Remark (Again on the identification between expected value of the indicator function and the probability of the corresponding set) Consider a probability space$(\Omega, \mathcal{F}, \mathrm{P})$, then if $A \subseteq \mathcal{F}$ is an event we can formally write

$$\mathrm{P}(A) = \int_A \mathrm{dP}(\omega); \tag{27}$$

and as a consequence

$$\mathbb{E}[I_A] = \int_\Omega I_A(\omega)\,\mathrm{dP}(\omega) = \int_A \mathrm{dP}(\omega) = \mathrm{P}(A).$$

In other words, the identification between the expected value of the indicator function and the probability of the event A becomes a trivial result. In the following, equation (27) will be used to define conditional expectations.

Based on the above considerations De Finetti suggested to use the same symbol of a set and for its indicator function and to use the same symbol for the probability of a set and the expectation of its indicator function. Here, these suggestions are not followed, but it is useful to be aware of them. However, the identification between expectations of indicator functions and probabilities of the corresponding sets can now be exploited to generalize a theorem which was presented in the previous section.

Theorem (Inclusion-exclusion formula) Let $A_1, \ldots, A_N$ be N events, then the probability of their union is given by

$$\mathrm{P}(\cup_{i=1}^N A_i) = \sum_{i=1}^N \mathrm{P}(A_i) - \sum_{i<j} \mathrm{P}(A_i \cap A_j) + \sum_{i<j<k} \mathrm{P}(A_i \cap A_j \cap A_k) \pm \ldots + (-1)^{N+1}\mathrm{P}(A_1 \cap \ldots \cap A_N). \quad (28)$$

Proof. Formula (28) is a direct consequence of the analogous formula for indicator functions. This can be derived by repeated application of the formula for the union of two sets using the associative property of unions:

$$\mathbb{I}_{\cup_{i=1}^N A_i} = \sum_{i=1}^N \mathbb{I}_{A_i} - \sum_{i<j} \mathbb{I}_{A_i}\mathbb{I}_{A_j} + \sum_{i<j<k} \mathbb{I}_{A_i}\mathbb{I}_{A_j}\mathbb{I}_{A_k} \pm \ldots + (-1)^{N+1}\mathbb{I}_{A_1}\ldots\mathbb{I}_{A_N}. \quad (29)$$

Taking the expectation of both sides and recalling its identification with the probability immediately leads to (28). □

A celebrated application of the inclusion-exclusion formula is to the so-called matching problem.

Example (The matching problem). At a dance N girls and their N boyfriends are paired at random. What is the probability of the event B "no girl dances with her boyfriend"? Consider the event A_i defined as "girl i dances with her boyfriend". The event $\cup_{i=1}^N A_i$ means "at least one girl dances with her boyfriend", therefore $\mathrm{P}(B) = 1 - \mathrm{P}(\cup_{i=1}^N)$. In order to apply equation (28) one first has to compute the probabilities $\mathrm{P}(A_i)$, $\mathrm{P}(A_i \cap A_j)$, etc.. Now, one has that $\mathrm{P}(A_i) = (N-1)!/N! = 1/N$. Indeed, there are $N!$ possible outcomes as girls as well as boyfriends can be numbered from 1 to N and there are $N!$ possible permutations of boyfriends. Now if girl i dances with her boyfriend, there are still $(N-1)!$ permutations for the other pairs. Similarly,

$\mathrm{P}(A_i \cap A_j) = (N-2)!/N! = 1/N(N-1)$. These results can also be justified as follows. The probability that girl i matches with her boyfriend is 1/N as there are N possible dance partners and only one is her boyfriend. The second result is a consequence of the definition of conditional probabilities

$$\mathrm{P}(A_i \cup A_j) = \mathrm{P}(A_i)\mathrm{P}(A_j|A_i)$$

and, as before, $PiA_i) = 1/N$ whereas $P(A_j|A_i) = 1/(N-1)$ as girl i is already paired with her boyfriend. Both lines of reasoning lead to

$$\mathrm{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \frac{1}{N(N-1)\ldots(N-k+1)}.$$

Notice that this result does not depend on the particular value of the indices, therefore, to apply (28), it is enough to count the number of terms in the sums on the right-hand side and they are $\binom{N}{k}$ when the intersection of k sets is considered. Now, one has that

$$\binom{N}{k} = \frac{N(N-1)\ldots(N-k+1)}{k!},$$

and finally

$$\begin{aligned}\mathrm{P}(\cup_{i=1}^{N}) = N\frac{1}{N} - \frac{N(N-1)}{2}\frac{1}{N(N-1)} \pm \ldots + (-1)^{N+1}\frac{1}{N!} \\ = 1 - \frac{1}{2!} + \frac{1}{3!} \pm \ldots + (-1)^{N+1}\frac{1}{N!}. \quad (30)\end{aligned}$$

If N is large enough the answer to the original question is

$$\mathrm{P}(B) = 1 - \mathrm{P}(\cup_{i=1}^{N}) \simeq 1 - \mathrm{e}^{-1}.$$

The expectation as well as the fundamental expectation theorem allow to define the moments of a distribution.

Definition (Moments of a distribution) Let X be a random variable and P_X its distribution, the k-th moment of the distribution ($k \in \mathbb{N}$) is

$$\mu_k = \mathbb{E}[X^k]. \quad (31)$$

If the expectation gives a typical value of a random variable X, the variance characterizes its deviations from this typical value.

Definition (Variance) The variance of a random variable X is

$$\mathrm{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]. \tag{32}$$

The variance is also called second central moment as it is the second moment of the random variable $Y = X - \mathbb{E}[X]$ centred around zero expectation. The following properties of the variance are immediate consequences of the definition. Their proof is left as an exercise to the reader.

Theorem (Properties of variance) If X is a random variable and $a \in \mathbb{R}$ one has

1. (multiplication by a constant)

$$\mathrm{var}(aX) = a^2 \mathrm{var}(X)$$

2. (variance and moments)

$$\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}^2[X] = \mu_2 - \mu_1^2.$$

Definition (Uncorrelated random variables) Two random variables X and Y defined on the same probability space are uncorrelated if

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \tag{33}$$

Theorem (Independence and absence of correlations) If X and Y are two independent random variables then they are uncorrelated.

Proof. The proof will be presented for two random variables with joint probability density function $p_{X,Y}(x,y) = p_X(x)p_Y(y)$. One has the following chain of equalities

$$\mathbb{E}[XY] = \int_{\mathbb{R}}\int_{\mathbb{R}} xy p_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathbb{R}}\int_{\mathbb{R}} xy p_X(x)p_Y(y)\,\mathrm{d}x\,\mathrm{d}y = \left(\int_{\mathbb{R}} x p_X(x)\,\mathrm{d}x\right)\left(\int_{\mathbb{R}} y p_Y(y)\,\mathrm{d}y\right) = \mathbb{E}[X]\mathbb{E}[Y],$$

thanks to theorems of Fubini-Tonelli type for multiple integrals. □

Notice that, in general, the converse is not true. Uncorrelated random variables are not independent.

Definition (Covariance) Given two random variables defined on the same probability space their covariance is

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]. \tag{34}$$

The following theorem is a consequence of the definition. The proof is left to the reader.

Theorem (Covariance) If X and Y are two random variables on the same probability space then

$$\operatorname{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

An immediate corollary if this theorem is that the covariance of two uncorrelated random variables vanishes.

Corollary (Covariance of uncorrelated random variables) Two random variables X and Y on the same probablity space are uncorrelated if and only if $\operatorname{cov}(X,Y) = 0$.

D. Conditional expectation

The concept of conditional expectation is very important in modern probability theory as it is used to define important classes of stochastic processes (random variables as a function of time).

Definition (Conditional probability space) Given a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and an event C, the conditional probability of any event A given C can be used to define a new probability space with P replaced by P_C, where for any $A \in \mathcal{F}$

$$\mathrm{P}_C(A) = \mathrm{P}(A|C) = \frac{\mathrm{P}(A \cap C)}{P(C)}.$$

The new probability space is $(\Omega, \mathcal{F}, \mathrm{P_C})$. Given a random variable X defined on the original probability space, it becomes natural to define a new random variable $X|C$ whose probability distribution $P_{X|C}$ is the image measure with respect to P_C; for any borel set $B \in \mathcal{B}(\mathbb{R})$ one has

$$P_{X|C}(B) = \mathrm{P}_C(X^{-1}B) = \frac{\mathrm{P}(X^{-1}B \cap C)}{\mathrm{P}(C)},$$

and one can define the conditional distribution function as well as the conditional probability density function

$$F_{X|C}(x) = P_{X|C}(X|C \leq x) = \frac{\mathrm{P}(X^{-1}[X \leq x] \cap C)}{\mathrm{P}(C)},$$

and

$$p_{X|C}(x) = \frac{\mathrm{d}F_{X|C}(x)}{\mathrm{d}x}.$$

Example (Conditional probability put into practice) Consider a random variable X on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and a Borel set $B \subseteq \mathcal{B}(\mathbb{R})$. Define $C = X^{-1}(B)$, then one can use the previous definitions to get

$$F_{X|C}(x) = F_X(x|B) = F_X(x|x \in B) = \frac{P_X[(X \leq x) \cap B]}{P_X(B)}$$

and the conditional cumulative probability distribution function can be written as a ratio of two integrals.

Definition (Conditional expectation) The expectation of a random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ has been defined via the image measure P_X as

$$\mathbb{E}[X] = \int_{\mathbb{R}} x P_X(\mathrm{d}x),$$

but one has that $P_X(\mathrm{d}x) = \mathrm{P}(X^{-1}\mathrm{d}x)$, let us denote by $\mathrm{d}\omega$ the inverse image of $\mathrm{d}x$ then

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\mathrm{P}(\mathrm{d}\omega).$$

In the same way, one can define the conditional expectation

$$\mathbb{E}[X|C] = \int_{\Omega} X(\omega)\mathrm{P}_C(\mathrm{d}\omega) = \int_{C} X(\omega)\mathrm{P}_C(\mathrm{d}\omega) = \frac{1}{\mathrm{P}(C)} \int_{C} X(\omega)\mathrm{P}(\mathrm{d}\omega).$$

Remark (Conditional expectation of the indicator function) Given an event A and its indicator function $\mathbb{I}_A$, not surprisingly one has

$$\mathbb{E}[\mathbb{I}_A|C] = \frac{1}{\mathrm{P}(C)} \int_{C} \mathbb{I}_A(\omega)\mathrm{P}(\mathrm{d}\omega) = \frac{1}{\mathrm{P}(C)}\mathrm{P}(A \cap C) = \mathrm{P}(A|C).$$

In other words, the conditional expectation of $\mathbb{I}_A$ given C coincides with the probability of A given C.

Definition (Conditional expectation with respect to a σ-field) Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and consider a partition of Ω into an exhaustive class of mutually disjoint sets C_i. Further consider the smallest σ-field $\mathcal{G} \subset \mathcal{F}$ generated by the partition. Let X be a random variable on the probability space and assign to each $\omega \in C_i$ the value $\mathbb{E}[X|C_i]$. This is a function from Ω to $\mathbb{R}$ and it is measurable with respect to the σ-field $\mathcal{G}$ as well as with respect to $\mathcal{F}$ which contains $\mathcal{G}$. Then this function is a random variable. It is denoted by the symbol $\mathbb{E}[X|\mathcal{G}]$ and can be defined as

$$\mathbb{E}[X|\mathcal{G}] = \sum_i \mathbb{E}[X|C_i]\mathbb{I}_{C_i}(\omega). \tag{35}$$

In this way, one also defines a new measure $\mathrm{P}_{\mathcal{G}}$ such that $\mathrm{P}_{\mathcal{G}}(G) = \mathrm{P}(G)$ if $G \subseteq \mathcal{G}$ and a corresponding probability space $(\Omega, \mathcal{G}, \mathrm{P}_{\mathcal{G}})$. Notice that any $G \subseteq \mathcal{G}$ can be written as $G = \cup_{i \in I} C_i$ for some set of indices I and one has

$$\mathrm{P}(G)\mathbb{E}[X|G] \;=\; \int_G X(\omega)\mathrm{P}(\mathrm{d}\omega) \;=\; \sum_{i \in I} \int_{C_i} X(\omega)\mathrm{P}(\mathrm{d}\omega) \;=\; \sum_{i \in I} \mathrm{P}(C_i)\mathbb{E}[X|C_i].$$

Indeed, using the definition of $(\Omega, \mathcal{G}, \mathrm{P}_{\mathcal{G}})$, one also has that

$$\mathrm{P}(G)\mathbb{E}[X|G] = \int_G X(\omega)\mathrm{P}(\mathrm{d}\omega) = \int_G \mathbb{E}[X|\mathcal{G}]\mathrm{P}_{\mathcal{G}}(\mathrm{d}\omega). \tag{36}$$

The latter chain of equalities means that from $\mathbb{E}[X|\mathcal{G}]$ by integration with respect to an event $G \subseteq \mathcal{G}$ one gets $\mathbb{E}[X|G]$. For this reason it is possible to call (35) the expectation conditioned to the σ-field $\mathcal{G}$ and not only the expectation conditioned to the partition. An important property of $\mathbb{E}[X|\mathcal{F}]$ is that its average coincides with the unconditional average of the random variable X, that is

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]. \tag{37}$$

This result is an immediate consequence of the definition (35). One has the following chain of equalities (recalling that $\{C_i\}$ is a partition of Ω):

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}\left[\sum_i \mathbb{E}[X|C_i]\mathbb{I}_{C_i}\right] = \sum_i \mathbb{E}[X|C_i]\mathbb{E}[\mathbb{I}_{C_i}] = \sum_i \mathbb{E}[X|C_i]\mathrm{P}(C_i) =$$
$$\frac{\mathrm{P}(C_i)}{\mathrm{P}(C_i)} \sum_i \int_{C_i} X(\omega)\mathrm{P}(\mathrm{d}\omega) = \sum_i \int_{C_i} X(\omega)\mathrm{P}(\mathrm{d}\omega) = \int_\Omega X(\omega)\mathrm{P}(\mathrm{d}\omega) = \mathbb{E}[X].$$

It is possible to extend the above construction to a generic σ-field $\mathcal{G} \subset \mathcal{F}$ by using equation (36) and defining $\mathbb{E}[X|\mathcal{G}]$ as the random variable $G(\omega)$ such that

$$\int_G G(\omega)\mathrm{P}_{\mathcal{G}}(\mathrm{d}\omega) = \int_G X(\omega)\mathrm{P}(\mathrm{d}\omega)$$

for any event $G \subseteq \mathcal{G}$. Based on a theorem due to Radon and Nikodym one can show that such a random variable exists and is uniquely determined except for a set of measure zero. Notice that, by definition, if X is $\mathcal{G}$ measurable then $\mathbb{E}[X|\mathcal{G}] = X$.

Example (Conditional expectation with respect to a random variable) Consider two random variables X and Y on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, let $\mathcal{X} \subset \mathcal{F}$ be the σ-field generated by the inverse images of Borel sets. One can now define the conditional average with respect to $\mathcal{X}$ as follows

$$\int_{X^{-1}B} \mathbb{E}[Y|\mathcal{X}]\mathrm{P}_{\mathcal{X}}(\mathrm{d}\omega) = \int_{X^{-1}B} Y(\omega)\mathrm{P}(\mathrm{d}\omega)$$

for any Borel set $B \subseteq \mathcal{B}(\mathbb{R})$. Now, it is possible to show that the function defined above is constant for any point $X^{-1}(a)$ for $a \in \mathbb{R}$ and, therefore, it is a function of the random variable X. The symbol $\mathbb{E}[Y|X]$ is normally used for the expectation of Y conditioned to X. One can also show that $\mathbb{E}[Y|X](\omega) = E[Y|X = a]$ for $\omega = X^{-1}(a)$.

The following theorem lists some important properties of conditional expectations which will be given without proof.

Theorem (Some properties of conditional expectation) Conditional expectation has the following properties

1. (Average of conditional expectation) $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$;
2. If X is $\mathcal{G}$-measurable then $\mathbb{E}[X|\mathcal{G}] = X$;
3. (Linearity) $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$;
4. (Positivity) If $X \geq 0$ then $\mathbb{E}[X|\mathcal{G}] \geq 0$;
5. (Tower property) If $\mathcal{H}$ is a sub-σ-field of $\mathcal{G}$ then $\mathrm{E}[\mathrm{E}[X|\mathcal{G}]|\mathcal{H}] = \mathrm{E}[X|\mathcal{H}]$;

6. (Taking out what is known) If Z is $\mathcal{G}$-measurable and bounded then $\mathbb{E}[ZX|\mathcal{G}] = Z\mathbb{E}[X|\mathcal{G}]$.

7. (Independence) If $\mathcal{H}$ is independent of $\sigma(\sigma(X), \mathcal{G})$ then $\mathbb{E}[X|\sigma(\mathcal{H}, \mathcal{G})] = \mathbb{E}[X|\mathcal{G}]$. In particular, if X is independent of $\mathcal{G}$ on has that $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.

Before ending this section, the reader must be warned that this introductory section on conditional probabilities and conditional expectation is not fully rigorous. The reference list contains textbook where the interested student can find a complete treatment of conditional expectations.

IV. INTRODUCTION TO STOCHASTIC PROCESSES

As mentioned in the previous section, in order to fully appreciate the importance of conditional expectation, one has first to define the concept of stochastic process. Unfortunately, it is impossible to reasonably discuss stochastic processes without reference to measure theoretic concepts. Here, stochastic processes in discrete time will be first introduced. Later, a non rigorous discussion of two important stochastic processes in continuous time will be presented. After reading this section you should be able to

- Define a stocastic process;
- Define martingales in discrete time and derive some of their properties;
- Define the simple random walk.

A. Finite dimensional distributions

Definition (Stochastic process) A family of ranom variables depending on the parameter $t \in T$, where T is an arbitrary set is called random function. One can write the symbol X_t or $X(t)$ to denote the family, but it is necessary to keep in mind that this is an application from $T \times \Omega$ to $\mathbb{R}$ and one should write $X(t, \omega)$. If T is a subset of real numbers and t is interpreted as time, then one uses the term *stochastic process* to denote the random function $X(t)$. If T is made up of integer numbers one can call the corresponding stochastic process as *discrete* stochastic process or as a *random sequence* or a *sequence of random variables.* A stochastic process is usually described in terms of finite dimensional distributions. For each k-tuple $(t_1, \ldots, t_k)$, of distinct elements of T, the random vector $(X(t_1), \ldots, X(t_k))$ has over $\mathbb{R}^k$ some joint finite dimensional distribution $\mu_{X(t_1),\ldots,X(t_k)}$

$$\mu_{X(t_1),\ldots,X(t_k)}(H) = P[(X(t_1), \ldots, X(t_k)) \in H] \tag{38}$$

for any Borel set $H \subseteq \mathcal{B}(\mathbb{R}^k)$. Two consistency conditions follow from (38). The first consistency condition is a symmetry requirement. Suppose that $H = H_1 \times \ldots \times H_k$ can be written as cartesian product of $H_i \in \mathcal{B}(\mathbb{R})$, as $(X(t_1), \ldots, X(t_k)) \in H_1 \times \ldots \times H_k$ this must be true also for any permutation $(\pi_1, \ldots, \pi_k)$ of the indices and $(X(t_{\pi_1}), \ldots, X(t_{\pi_k})) \in H_{\pi_1} \times \ldots \times H_{\pi_k}$ is the same event. Therefore, for any permutation of the indices one must have that

$$\mu_{X(t_1),\ldots,X(t_k)}(H_1 \times \ldots \times H_k) = \mu_{X(t_{\pi_1}),\ldots,X(t_{\pi_k})}(H_{\pi_1} \times \ldots \times H_{\pi_k}). \tag{39}$$

The second consistency condition is the relationship between $\mu_{X(t_1),\ldots,X(t_{k-1})}$ and $\mu_{X(t_1),\ldots,X(t_k)}$:

$$\mu_{X(t_1),\ldots,X(t_{k-1})}(H_1 \times \ldots \times H_{k-1}) = \mu_{X(t_1),\ldots,X(t_k)}(H_1 \times \ldots \times H_{k-1} \times \mathbb{R}); \tag{40}$$

indeed, $(X(t_1), \ldots, X(t_{k-1})) \in H_1 \times \ldots \times H_{k-1}$ if and only if $(X(t_1), \ldots, X(t_k)) \in H_1 \times \ldots \times H_{k-1} \times \mathbb{R}$.

Finite dimensional distributions coming from a stochastic process obey the two consistency conditions (39) and (40). Conversely, by means of Kolmogorov's constructive existence theorem, one can show that given measures satisfying the two consistency conditions, there exists a stochastic process having these measures as finite dimensional distributions. Kolmogorov's existence theorem can be used to prove the existence of a given stochastic process, if one can write explicitly the finite dimensional distributions and prove that they satisfy the two consistency conditions.

B. Discrete stochastic processes

Let now T coincide with the set of natural numbers $\mathbb{N}$. The value $X(t = n, \omega)$ of the stocastic process at time (or step) $t = n$ will be denoted by $X_n(\omega)$ or, simply, by X_n.

Definition (Filtered space) A filtered probability space is a quadruple $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathrm{P})$ where $(\Omega, \mathcal{F}, \mathrm{P})$ is a usual probability space and $\{\mathcal{F}_n : n \geq 0\}$ is a filtration - an increasing family of sub-σ-fields of $\mathcal{F}$:

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}.$$

The σ-field $\mathcal{F}_\infty$ is defined as

$$\mathcal{F}_\infty = \sigma\left(\cup_{i=0}^{\infty} \mathcal{F}_i\right) \subseteq \mathcal{F}.$$

Usually $\{\mathcal{F}_n\}$ is the *natural* filtration, that is $\mathcal{F}_n$ is the σ-field generated by the random variables $X_0, \ldots, X_n$: $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$ and the information about ω that one has at step n (or better, just after step n) consists of the values $X_0(\omega), \ldots, X_n(\omega)$.

Definition (Adapted process) A process $X = (X_n : n \geq 0)$ is called adapted to the filtration $\{\mathcal{F}_n\}$ if, for each n, X_n is $\mathcal{F}_n$-measurable. Recalling property 2 of conditional averages this means that $\mathbb{E}[X_n|\mathcal{F}_n] = X_n$ or, in other words, that the value of X_n is known at step n.

Definition (Martingale, submartingale, supermartingale) A process $X = (X_n : n \geq 0)$ on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathrm{P})$ is called a martingale if

1. X is adapted;
2. $\mathbb{E}[|X_n|] < \infty \; \forall n$;
3. $\mathbb{E}[X_n|\mathcal{F}_{n-1}] = X_{n-1}$ $(n \geq 1)$.

The martingale is a rigorous statement of the intuitive concept of fair game. If, the first two properties are satisfied, but $\mathbb{E}[X_n|\mathcal{F}_{n-1}] \leq X_{n-1}$ for $n \geq 1$, one has a supermartingale and if $\mathbb{E}[X_n|\mathcal{F}_{n-1}] \geq X_{n-1}$ for $n \geq 1$, one has a submartingale. If the random variable X_0 is $\mathcal{L}^1(\Omega, \mathcal{F}_0, \mathrm{P})$, then X is a martingale (repectively supermartingale, submartingale) if and only if $(X - X_0 : X_n - X_0, n \in \mathbb{Z}^+)$ is a martingale (supermartingale, submartingale). For this reason, it is sufficient to study martingales (supermartingales, submartingales) with $X_0 = 0$.

Theorem (Consequence of the tower property) If $X = (X_n : n \geq 0)$ is a martingale then $\mathbb{E}[X_n|\mathcal{F}_m] = X_m$ for $m < n-1$.

Proof. $\mathcal{F}_m \subseteq \mathcal{F}_i$ is a sub-σ-field of $\mathcal{F}_i$ for $m+1 \leq i \leq n-1$ therefore repeated applications of the tower properties show that

$$\mathrm{E}[X_n|\mathcal{F}_m] = \mathrm{E}[X_n|\mathcal{F}_{n-1}|\mathcal{F}_m] = \mathrm{E}[X_{n-1}|\mathcal{F}_m] = \ldots \mathrm{E}[X_{m+1}|\mathcal{F}_m] = X_m.$$

□

Example (Sum of independent random variables with zero mean: Random walk) Consider a sequence of independent random variables $X_1, X_2, \ldots$ with $\mathbb{E}(|X_k|) < \infty$, $\forall k$ and $\mathbb{E}(X_k) = 0$, $\forall k$. Define

$$S_n = \sum_{k=1}^{n} X_k,\ S_0 = 0; \tag{41}$$

The stochastic process S is called a random walk and it has independent increments by construction. Let us define the filtration $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, $\mathcal{F}_0 = (\emptyset, \Omega)$. Then, for $n \geq 1$, one has

$$\mathbb{E}[S_n|\mathcal{F}_{n-1}] = \mathbb{E}[S_{n-1}|\mathcal{F}_{n-1}] + \mathbb{E}[X_n|\mathcal{F}_{n-1}] = S_{n-1} + \mathbb{E}[X_n] = S_{n-1},$$

where the first equality is a consequence of the linearity of the conditional expectation, and the second equality uses property 2 and property 7 (independence); finally, the last equality is a consequance of the definition of S. In other words, the above chain of equalities shows that the sum of independent random variables with zero mean is a martingale.

Example (Product of non-negative independent random variables with mean 1) Consider a sequence of independent non-negative random variables $X_1, X_2, \ldots$ with $\mathbb{E}(X_k) = 1$, $\forall k$. Let us consider the same filtration as in the previous example and the process

$$M_n = X_1 X_2 \ldots X_n,\ M_0 = 1;$$

then for $n \geq 1$, one has

$$\mathbb{E}[M_n|\mathcal{F}_{n-1}] = \mathbb{E}[M_{n-1}X_n|\mathcal{F}_{n-1}] = M_{n-1}\mathbb{E}[X_n] = M_{n-1},$$

where the first equality is based on the definition of the process, the second is a consequence of property 6 (taking out what is known) and property 7 (independence); finally, the third and last equality again follows from the definition of the process. In other words, the above chain of equalities shows that the product of independent random variables with mean 1 is a martingale.

These two examples are quite important both in finance and in economic theory. The random walk is used as a simple model of price fluctuations in financial markets. This idea was introduced by L. Bachelier in his PhD thesis published in 1900. Multiplicative models have been used to describe the growth of firms at least since Gibrat's work published in the 1930's.

A process with independent increments with zero mean is a martingale. The converse is not true; however, one can prove that a martingale has uncorrelated increments.

Theorem (A martingale has uncorrelated increments) Let the process S be a martingale with respect to the filtration $\{\mathcal{F}_n\}$, then S has uncorrelated increments.

Proof. Consider the increments $X_n = S_n - S_{n-1}$ and $X_{n+1} = S_{n+1} - S_n$, one has to show that $\mathbb{E}[X_{n+1}X_n] = 0$. Indeed, one has the following chain of equalities

$$\mathbb{E}[X_{n+1}X_n] = \mathbb{E}[(S_{n+1}-S_n)(S_n-S_{n-1})] = \mathbb{E}[\mathbb{E}[(S_{n+1}-S_n)(S_n-S_{n-1})]|\mathcal{F}_{n-1}]] = \mathbb{E}[(S_{n+1} - S_n)\mathbb{E}[(S_n - S_{n-1})|\mathcal{F}_{n-1}]] = 0,$$

where the second equality is a consequence of the fact that $(S_{n+1}-S_n)(S_n-S_{n-1})$ is $\mathcal{F}_{n-1}$-measurable, the third equality uses the property "taking out what is known" and the last equality uses the fact that S is a martingale and that the expected value of a constant (0 in this case) is the constant itself. □

Example (The efficient market hypothesis) In discrete time, the efficient market hypothesis can be stated as a (sub)martingale hypothesis on the price process. Suppose there is a market with two assets: a risky asset (a share paying dividends) and a risk-free asset (a zero-coupon bond). let $r_A(t, t+1)$ denote the return of the risky asset in a period and $r_F(t, t+1)$ denote the return of the risk-free asset in the same period. At the beginning of the period the return of the risk-free asset is known, whereas the return of the risky asset is

$$r_A(t, t+1) = \frac{P_A(t+1) + D_A(t, t+1) - P_A(t)}{P_A(t)},$$

where $P_A(t)$ represents the price of the risky asset at time t and $D_A(t, t+1)$ is the dividend paid by the risky asset A in the period, but it is not known at the beginning of the period as $P_A(t)$ as well as $D_A(t, t+1)$ are stochastic processes. However, given the information available at time t, the beginning of the period, if the expected return of the risky asset were systematically higher that the return of the risk-free asset, in principle one could borrow money at the risk-free rate and use it to invest in the risky asset in order to get a gain at the end of the period. Analogously, if the expected return of the risky asset were systematically lower than the risk free return, in principle one could sell short the risky asset, invest the money in the riskless asset in order to get a gain at the end of the period. Both these schemes would violate the principle of no-arbitrage that in a market there is "no free lunch". Therefore, one is led to impose that the expected return on the risky asset be equal to the risk-free return:

$$\mathbb{E}[r_A(t, t+1)|I(t)] = r_F(t, t+1), \tag{42}$$

where $I(t)$ represents the information available at the beginning of the period.

In order to better understand the implications of equation (42) on the price process $P_A(t)$, assume that the risky asset pays no dividend in the period and that the risk-free rate is a constant, r_F, then one gets (if $r_F > 0$):

$$\mathbb{E}[P_A(t+1)|I(t)] = (1 + r_F)P_A(t) \geq P_A(t),$$

and (if $r_F = 0$)

$$\mathbb{E}[P_A(t+1)|I(t)] = P_A(t);$$

this means that either the price process is a submartingale or it is a martingale with respect to the filtration represented by $I(t)$. Notice that, in a real market, $I(t)$ is usually larger than the natural filtration used in probability theory which is only based on the past history of the process. A consequence of the martingale property is that serial correlations between increments of the price process as well as returns are ruled out. Empirical analyses on stock shares or stock indices fluctuations show that this is the case. However, uncorrelated increments does not mean that the increments are independent and, indeed, independence is also ruled out by empirical analyses.

V. MARKOV CHAINS

The purpose of this section is

VI. POISSON PROCESS

The purpose of this section is

VII. WIENER PROCESS

The purpose of this section is

Appendix A: Some important distributions

The purpose of this section is to present some important distributions that are widely used in the applications. The main focus is on discrete distributions. Every complete introductory textbook on probability theory contains a list of important distributions and their main properties.

Discrete distributions

Consider a countable sample space Ω and use $\mathcal{F} = \mathcal{P}\Omega$ as its σ-field as done above in the remark on probabilities on a countable space. Now, given a random variable X on a countable sample space, its image $T' = \{X(\omega) : \omega \in \Omega\}$ is either finite or countably infinite and the usual definition of distribution

$$P_X(A) = \mathrm{P}(\omega : X(\omega) \in A)) = \mathrm{P}(X^{-1}A) = \mathrm{P}(X \in A)$$

defines a probability measure on T' equipped with the σ-field $\mathcal{P}T'$ of all subsets of T'. T' is at most countable and, therefore, the distribution is completely determined by the numbers

$$p_{X,j} = \mathrm{P}(X = j) = \sum_{\{\omega_i\colon X(\omega_i)=j\}} p_i.$$

The family $(p_{X,j} : j \in T')$ is also called the distribution or the law of X. One has that $P_X(A) = \sum_{j\in A} p_{X,j}$. If P_X is a known distribution, for instance the Poisson distribution, one can say that X is a Poisson random variable.

Remark (Expected value for a discrete random variable) Let X be a random variable on a countable probability space $(\Omega, \mathcal{F} = \mathcal{P}\Omega, \mathrm{P})$, then the expected value of X is given by

$$\mathbb{E}[X] = \sum_i X(\omega_i)p_i = \sum_{j\in T'} jp_{X,j} = \sum_{j\in T'} j\mathrm{P}(X = j), \tag{43}$$

provided either the series is absolutely convergent or $X \geq 0$. In the latter case one can also have $\mathbb{E}[X] = +\infty$.

An important class of inequalities follow from the next theorem.

Theorem (Basic inequality) Let $h : \mathbb{R} \to [0, \infty)$ be a non-negative function and let X be a random variable. Then

$$\mathrm{P}(\{\omega : h(X(\omega)) \geq a\}) \leq \frac{\mathbb{E}[h(X)]}{a}, \tag{44}$$

for any positive constant a.

Proof. Let $Y = h(X)$ and consider the set

$$A = \{Y^{-1}[a, \infty)\} = \{\omega : h(X(\omega)) \geq a\} = \{h(X) \geq a\};$$

then $h(X) \geq a\mathbb{I}_A$. Now, given two positive random variables Y and Z on the same probability space and such that $Y \geq Z$, using the properties of integrals, one can show that $\mathbb{E}[Y] \geq \mathbb{E}[Z]$. Using $Y = h(X)$ and $Z = a\mathbb{I}_A$ one has

$$\mathbb{E}[h(X)] \geq \mathbb{E}[a\mathbb{I}_A] = a\mathbb{E}[\mathbb{I}_A] = a\mathrm{P}(A) = a\mathrm{P}(\{\omega : h(X(\omega)) \geq a\}).$$

□

This theorem has the so-called Markov's inequality as its immediate corollary, where $h(X) = |X|$:

Corollary (Markov's inequality) If X is a random variable, then

$$\mathrm{P}(\{|X| \geq a\}) \leq \frac{\mathbb{E}[|X|]}{a}. \tag{45}$$

Corollary (Chebyshev's inequality 1) Consider a random variable X, then for any positive constant a one has

$$\mathrm{P}(\{|X| \geq a\}) \leq \frac{\mathbb{E}[X^2]}{a^2}. \tag{46}$$

Proof. Also this inequality follows from (44). It is sufficient to notice that the set $\{|X| \geq a\}$ coincides with the set $\{X^2 \geq a^2\}$ and use $h(X) = X^2$:

$$\mathrm{P}(\{|X| \geq a\}) = \mathrm{P}(\{X^2 \geq a^2\}) \leq \frac{\mathbb{E}[X^2]}{a^2}.$$

□

Corollary (Chebyshev's inequality 2) If $Y = |X - \mathbb{E}[X]|$ replaces X in (46), one gets

$$\mathrm{P}(\{|X - \mathbb{E}[X]| \geq a\}) \leq \frac{\mathrm{var}(X)}{a^2}; \tag{47}$$

this inequality is also called Bienaymé-Chebyshev inequality.

Coming back to discrete distributions, in principle any sequence of non-negative terms summing to 1 is allowed. In practice, there are some sequences that recur in the description of natural phenomena. They will be described in the following.

Example (The Poisson distribution) The Poisson distribution of parameter $\lambda > 0$ is defined as

$$p_{X,n} = \mathrm{P}(X = n) = \mathrm{e}^{-\lambda}\frac{\lambda^n}{n!};\ n \geq 0 \tag{48}$$

The expected value is λ:

$$\mathbb{E}[X] = \sum_{j=0}^{\infty} j\mathrm{P}(X = j) = \sum_{j=0}^{\infty} j\frac{\lambda^j}{j!}\mathrm{e}^{-\lambda} = \lambda \sum_{j=1}^{\infty} \frac{\lambda^{j-1}}{(j-1)!}\mathrm{e}^{-\lambda} = \lambda\mathrm{e}^{\lambda}\mathrm{e}^{-\lambda} = \lambda.$$

Example (The Bernoulli distribution) X has the Bernoulli distribution if X assumes only two values 0 and 1. Usually $\{X = 1\}$ corresponds to success and $\mathrm{P}(\{X = 1\}) = p$ whereas $\{X = 0\}$ corresponds to failure and $\mathrm{P}(\{X = 0\}) = q = 1-p$. The expected value of X is $\mathbb{E}[X] = 1 \cdot p + (1-p) \cdot 0 = p$. The variance of X is given by $\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}^2[X] = p - p^2 = p(1-p) = pq$.

Example (Binomial distribution) X has the binomial distribution if, for given n, X can only assume the values $\{0, 1, \ldots, n\}$ with

$$\mathrm{P}(\{X = k\}) = \binom{n}{k} p^k (1-p)^{n-k} \tag{49}$$

where $p \in [0, 1]$ and n are the two parameters of the distribution. Suppose one performs n times an experiment following Bernoulli's distribution and let $Y_1, \ldots, Y_n$ be the corresponding Bernoulli random variables. The random walk $X = \sum_{i=1}^{n} Y_i$ has the binomial distribution and

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathbb{E}[Y_i] = np.$$

The variance of X can be computed by noting that the Y_i are independent and identically distributed random variables. Therefore, $\mathbb{E}[Y_i Y_j] = 0$ for $i \neq j$. Therefore one has

$$\mathrm{var}(X) = n\,\mathrm{var}(Y_i) = npq.$$

Both results can be also directly obtained by computing sums such as

$$\mathbb{E}[X] = \sum_{i=0}^{n} i\mathrm{P}(X = i) = \sum_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i},$$

but this is more painful.

Example (The hypergeometric distribution) A typical model for the binomial distribution is called sampling with replacement from an urn. Suppose that in a box there are r red balls and b black balls and that a success corresponds to extracting a black ball. After an extraction, the color of the ball is recorded and then it is returned to the box, the box is mixed and a new ball is extracted and so on. If this experiment is repeated n times, $p = b/(r+b)$, X is the number of successes, one shows that $\mathrm{P}(X = j)$ follows the binomial distribution.

If the same experiment is performed without replacing the ball, then one gets the hypergeometric distribution. Again consider an urn with r red balls and b black balls and let $N = r + b$ be the total number of balls. Let n be the number of trials or samples taken from the urn. Then the probability of getting j black balls out of n can be computed according to the classical definition of probability. n balls are drawn without replacement from the box. The total number of possible cases is given by the number of possible choices of n objects out of N. This is the binomial coefficient $\begin{pmatrix} N \\ n \end{pmatrix}$. Again, success is a black ball. If the total number of favourable cases is j, one has first to select j black balls out of b and then $n-j$ red balls out of $N-b$. Therefore the number of favourable cases is $\begin{pmatrix} b \\ j \end{pmatrix}\begin{pmatrix} N-b \\ n-j \end{pmatrix}$. As a consequence, a random variable has the hypergeometric distribution of parameters (N, b, n) if

$$\mathrm{P}(X = j) = \frac{\begin{pmatrix} b \\ j \end{pmatrix}\begin{pmatrix} N-b \\ n-j \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}, \tag{50}$$

for $0 < n < N$. If N and b are large the hypergeometric distribution can be approximated by the binomial distribution of parameter $p = b/N$ when the total number of trials is small, $n << N$. Indeed, in this case, removing a few balls from the urn does not change much its composition. The expected value of a hypergeometric random variable is $\mathbb{E}[X] = nb/N = np$. Its variance is $\mathrm{var}(X) = n\,p(1-p)\frac{N-n}{N-1}$.

Example (Geometric distribution) Let us consider a sequence of independent Bernoulli trials. Instead of fixing the number of trials n, one is now interested in achieving a certain given number of successes. If this number is one, the random variable X denoting the number of trials needed to have one success follows the so-called geometric distribution

$$\mathrm{P}(X = j) = (1-p)^{j-1}p. \tag{51}$$

The expected value of a random variable following the geometric distribution is

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} j\mathrm{P}(X = j) = \sum_{J=1}^{\infty} j(1-p)^{j-1}p = p\sum_{j=1}^{\infty} jq^{j-1} = p\frac{\mathrm{d}}{\mathrm{d}q}\left(\frac{1}{1-q}\right) = \frac{1}{p}.$$

The variance is $\mathrm{var}(X) = (1-p)/p^2$.

Example (Negative binomial distribution) Now suppose to go on with Bernoulli trials until the r-th success takes place. If one consider a single sequence of these trials, with r successes and j failures, its probability is $p^r(1-p)^j$. The number of such sequences can be determined by observing that the last element of each sequence must be a success. Then, one has to choose j failures out of $r+j-1$ positions in the sequence and if the failures define the random variable X

$$\mathrm{P}(X=j) = \binom{j+r-1}{j} p^r (1-p)^j. \tag{52}$$

This is the Pascal (or negative binomial) distribution. If one is interested in the total number of trials necessary to see r successes, then one can define a new random variable $Y = X + r$. As the total number of trials to see one success follows the geometric distribution, and trials are independent, the total number of trials to see r successes is a sum of r independent and identically distributed geometric random variables, $\{Z_i\}_{i=1}^r$, of parameter p: $Y = \sum_{i=1}^r Z_i$. One has that $\mathbb{E}[Y] = r/p$ and $\mathbb{E}[X] = \mathbb{E}[Y] - r = r(1-p)/p$.

Continuous distributions

Appendix B: Elements of measure theory

Include a discussion of dominated convergence.

---

[1] A. Aczel, *Chance: A Guide to Gambling, Love, the Stock Market, and Just About Everything Else*, Italian translation *Chance*, Raffaello Cortina Editore, Milano, 2005.
This is a popular-science book written by a mathematician. It includes tutorial descriptions of many famous problems of probability theory.

[2] P. Billingsley, *Probability and Measure*, Third Edition, Wiley, New York, 1995.
This is a standard reference for contemporary measure-theoretic probability theory. It includes a discussion of Kolmogorov's existence theorem and its application to the Wiener process.

[3] N. Bouleau, *Processus stochastiques et applications*, Hermann, Paris, 2000.

A nice introductory textbook on stochastic processes written for engineers in the rigorous tradition of French math texbooks. It is available only in French.

[4] C. Conti, *Probabilità e valore nelle scienze sociali*, Mazzotta Editore, Milano, 1975.

This book, available only in Italian, discusses the history of probabilistic concepts in the social sciences with a focus on Economics and on the theory of value.

[5] D. Costantini, *Fondamenti del calcolo delle probabilità*, Feltrinelli Editore, Milano, 1970.

This book is available only in Italian. It contains a historical account of the foundations of probability theory.

[6] R. Durrett, *The Essential of Probability*, The Duxbury Press, Belmont California, 1994.

A no-nonsense introduction to rigorous probability theory without much measure theory. This is a very good introductory textbook for a one semester course.

[7] I. Hacking, *The Emergence of Probability*, Second Edition, Cambridge University Press, Cambridge, 2006.

It is a classical book on the early history of probability theory. First published in 1975, it triggered many other studies. The subtitle *A Philosophical Study of Early Ideas About Probability Induction and Statistical Inference* shows that Hacking did not have in mind to write just a historical essay.

[8] I. Hacking, *An Introduction to Probability and Inductive Logic*, Cambridge University Press, Cambridge, 2001.

This is a textbook on probability written for students of Philosophy.

[9] G. Kersting and A. Wakolbinger, *Elementare Stochastik*, Birkhäuser Verlag, Basel, 2008.

A recent textbook written for beginners and for a short course on probability theory. It does not include measure-theoretic methods.

[10] J. Jacod and P. Protter, *Probability Essentials*, Springer Verlag, Berlin, 1991.

A no-nonsense textbook with short chapters containing essential information on many aspects of probability theory.

[11] C.M. Monti and G. Pierobon, *Teoria della probabilità*, Decibel Editore, Padova and Zanichelli Editore, Bologna, 2000.

This is a textbook for Italian engineers. It contains many exercises and examples.

[12] D. Pollard, *A User's Guide to Measure Theoretic Probability*, Cambridge University Press, 2002.

This is an advanced textbook on measure-theoretic probability. Read this first and Billingsley's treatise will have no more misteries for you.

[13] R. Schilling, *Measures, Integrals and Martingales*, Cambrige University Press, Cambridge, 2005.

Everything you need on measure theory and more! A contemporary treatise you can read before or together with Billingsley's book.

[14] G. Shafer and V. Vovk, *Probability and Finance. It's only a game*, Wiley, New York, 2001.

If you forget about the title, you will have in your hands a book where the authors try to reconstruct probability theory based on a game-theoretic approach. The approach mediates between the frequentist and the subjectivist points of view and many classical results are recovered in the game-theoretic framework. Chapter 2 (on the historical context) is very interesting.

[15] M.R. Spiegel, *Real Variables*, McGraw Hill, New York, 1969.

A classical elementary introduction to measure theory, Lebesgue integration and Fourier series.

[16] D.W. Stroock, *An introduction to Markov Processes*, Springer Verlag, Berlin, 2000.

There are plenty of books devoted to Markov processes. Stroock himself is fond of the book by Karlin and Taylor (S. Karlin and H. Taylor, *A First Course in Stochastic Processes*, Academic Press, New York, 1975). Stroock's book contains a review of results on Markov processes and discusses reversible Markov processes and their application to Monte Carlo simulations.

[17] N.N. Taleb, *The Black Swan*, Penguin Books, London, 2008.

This is not the best, but the most successful of the writings by N.N. Taleb. The book contains a criticism of classical probability theory as well as of the way in which it is used in mathematical finance. According to Taleb, probability theory deals with a sort of tamed, *predictable* chance, whereas the events in human history are essentially unpredictable.

[18] A.D. Ventsel, *Kurs teorii sluciajnych prozessov*, Italian translation *Teoria dei processi stocastici*, Editori Riuniti, Roma, 1983.

This is a book of the Russian school on stochastic processes. It contains an overview of many methods used in the theory. The book was written based on the lectures made by the author in 1969 at Moscow University.

[19] D. Williams, *Probability with Martingales*, Cambridge University Press, Cambridge, 1991.

This textbook is a recommended introduction to measure-theoretic probability. Even if it has

introductory character, it contains rigorous proofs of the strong law of large numbers, a topic usually discussed only in advanced textbooks or monographs. Moreover, here, the student can learn most of martingale theory for discrete stochastic processes and then read one of the good treatises on stochastic integrals with profit.