BASIC PROBABILITY THEORY 2024

JUHAN ARU

[1]

[1]Version of 2024. All kinds of feedback, including smaller or bigger typos, is appreciated - juhan.aru@epfl.ch. This is a third version of the notes. In writing previous version of these notes I have consulted notes of I. Manolescu (Fribourg), Y. Velenik (Geneva), A. Eberle (Bonn) (all on their websites) and the book by R. Dalang & D. Conus published by EPFL press.

Section 0

Introduction

This course is about probability theory: the mathematical framework for formalising our questions about random phenomena, and their mathematical study.

When we want to describe a random phenomena in the real world, we build a mathematical model. This is itself an interesting process and a good model involves lots of well-chosen simplifications and righteous choices - e.g. to model a coin toss, we usually discard the possibility of it landing on the edge, or without further knowledge we consider the heads and tails equiprobable, although that may not be the case for example already because of different weight distributions. But this all is not the topic of this course.

In this course we will study the general mathematical framework and formulation of such models and then discuss the mathematical tools necessary and useful to study such models. Hopefully we also have some time to discuss some interesting models.

Section 1

Basic framework

In this chapter we discuss some basic but important notions of probability theory:

- Probability space
- Random variables
- Independence

1.1 Probability space

Our first aim is to motivate the notion of a probability space or a probabilistic model. To do this let us consider two examples:

(1) A random number with values in $\{1, 2, \ldots, 12\}$ e.g. something that comes from a lottery.
(2) Describing the weather in Lausanne the day after.

In describing these two random phenomena we will still use everyday vocabulary / intuitions. Thereafter we will give the mathematical definitions that will fix the vocabulary for the rest of the course.

(1) Random number. To describe a random number mathematically, we basically need three inputs:

- The set of all possible outcomes: in this case $\Omega = \{1, 2, 3, \ldots, 12\}$
- The collection of yes / no questions that we can answer about the actual outcome, i.e. this random number. For example:
  - Is this number equal to 3?
  - Is this number even?
  - Is this number smaller than 4?

  To each of these questions we put in correspondence the subset of outcomes that corresponds to the answer yes: $\{3\}$, $\{2, 4, 6, 8, 10, 12\}$ or $\{1, 2, 3\}$ respectively. We call each such subset an event.
- Finally, to each event $E \subseteq \Omega$ we want to assign a numerical value $\mathbb{P}(E) \in [0, 1]$ that we call the probability. This should correspond to the fraction of times an event happens if the random number is given to us many times, e.g. if the lottery is played many times. [2]

Here the set of possible outcomes was easy and directly given by the problem. Also it is natural to assume that each subset $E \subseteq \Omega$ is an event - or in other words that for each E we can ask the question: is the number in E? This means that the we can take the collection of events to correspond to all subsets of Ω.

Determining the probability really depends on what we want to model - e.g. if we are trying to model the lottery, we may assume that all numbers are equally likely and then we

[2] In fact, one uses probabilistic models also to model phenomena that only happens once. In that case probability measures somehow our degree of belief.

rediscover the model from high-school: we set $\mathbb{P}(E) = |E|/|\Omega|$. However, if we wanted to describe the sum of two dice, we would need to choose the numbers $\mathbb{P}(E)$ very differently! [3]

Now, if we want our model to correspond to the intuitive notion of probability and to predict the fraction of repeated experiments, then these choices are not quite free - we need to add some constraints. E.g. we cannot put in an arbitrary function $\mathbb{P}$: indeed, if we have two events $E_1 \subseteq E_2$ then we should have $\mathbb{P}(E_1) \leq \mathbb{P}(E_2)$ as every time E_1 happens, also E_2 happens. We should also have $\mathbb{P}(\Omega) = 1$ as something always happens and $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ if E and F are disjoint (why?). Of course not all these constraints are distinct - some might imply others and when giving the definition of a probability space below we will purify and choose only some conditions that will then mathematically imply all the others.

(2) Weather in Lausanne the day after. We would again want to make the three decisions, but here the task is already harder at the very first step. What should be the state space? A natural state space could probably be all possible microscopic states of the atmosphere up to 20km of height over Lausanne...but here we of course have many arbitrary choices - why 20 km, how wide should we look over Leman etc? And in any case, any natural state would be impossibly complicated!

Luckily, we do not actually need to worry about it - we only have to assign probabilities to all the events in the collection of events. And we have some freedom in choosing this collection events - it could be determined by our possibility to measure the states, e.g. we are able to measure the temperature up to some precision, or the density of CO_2 or water molecules to some precision and this determines some subsets of the state space.

However, as with the probability function, also for the collection of events there are some natural consistency conditions: we would assume that if one can observe if event E happened, we should be also able to measure if its complement E^c happened. Or if we are able to say if E happened or if F happened, we should be able to say if one of the two happened - i.e. $E \cup F$ should also be an event. And in fact it comes out that this is all we need!

Naturally, setting up probabilities for this model is also horribly complicated - there are no natural symmetry assumptions like the one we used for the uniform distribution. Also, even the best physicist in the world will not be able to describe the natural probability distribution of all microscopic states of the atmosphere, especially as it will heavily depend on what is happening just before! Thus, our only choice basically is to try to somehow use the combination of our knowledge about atmospheric processes together with our observations from history to set up some estimates for the model; and then naturally we will try to improve it with every next day. Luckily, this difficult task is not up to us but rather the office of meteo and the statisticians!

Remark 1.1. *Finally, before giving the mathematical definitions, let us stress again that all three components of the model - the sample space, the set of events and their probabilities - are inputs that we choose to build our model. When trying to model a real world phenomena we usually make simplifications for each of these choices. For example, for the coin toss we use only two outcomes: heads and tails, although theoretically edge is also possible. Also, we usually set probabilities to be a half, although that is not exactly true either.*

[3] See Exercise sheet 1.

1.2 Mathematical definition of a probability space

We are now ready to use our mathematical filter and give a mathematical definition of a probability space. In fact, we first use the mathematical purifier to come up with a definition in the restricted setting where Ω is a finite set, and then generalize it further.

Indeed, the discussions above lead us directly to:

Definition 1.2 (Elementary probability space, Kolmogorov 1933)**.** *An elementary probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where*

- *Ω is a finite set, called the state or sample space or the universe.*
- *$\mathcal{F}$ is a set of subsets of Ω, satisfying:*
  - *$\emptyset \in \mathcal{F}$;*
  - *if $A \in \mathcal{F}$, then also $A^c \in \mathcal{F}$;*
  - *If $A_1, A_2, \in \mathcal{F}$, then also $A_1 \cup A_2 \in \mathcal{F}$.*

  $\mathcal{F}$ is called the collection of events and any $A \in \mathcal{F}$ is called an event.
- *And finally, we have a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying $\mathbb{P}(\Omega) = 1$ and additivity for disjoint sets: if $A_1, A_2 \in \mathcal{F}$ are pairwise disjoint, then*

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2).$$

  This function $\mathbb{P}$ is called the probability

Notice that some properties discussed above, like the fact that for events $E_1 \subseteq E_2$, we have $\mathbb{P}(E_1) \leq \mathbb{P}(E_2)$, follow directly from the definition.[4]

Now, most phenomena in the real world can be described by finite sets just because we are able to measure things only to a finite level of precision. However, like the notion of a continuous or differentiable function helps to simplify our mathematical descriptions of reality and thus improve our understanding, continuous probability spaces also make the mathematical descriptions neater, simpler and thereby also make it easier to understand and study the underlying random phenomena.

Some natural examples where infinite sample spaces come in: an uniform point on a line segment e.g. stemming from breaking a stick into several pieces; the position on the street where the first raindrop of the day falls; or the space of all infinite sequences of coin tosses. In all these cases the mathematically natural state space is even uncountable. Countably infinite state spaces can also come up: for example if we want to model the first moment that a repeated coin toss comes up heads, the value might be $1, 2, 3$ or with very very small probability also 10^{10}, so a natural state space would contain all natural numbers.

So let us state the general definition:

Definition 1.3 (Probability space, Kolmogorov 1933)**.** *A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where*

- *Ω is a set, called the state or sample space or the universe.*
- *$\mathcal{F}$ is a set of subsets of Ω, satisfying:*
  - *$\emptyset \in \mathcal{F}$;*
  - *if $A \in \mathcal{F}$, then also $A^c \in \mathcal{F}$;*
  - *If $A_1, A_2, \cdots \in \mathcal{F}$, then also $\bigcup_{n\geq 1} A_n \in \mathcal{F}$.*

  $\mathcal{F}$ is called the collection of events or a σ-algebra and any $A \in \mathcal{F}$ is called an event.

[4] See Exercise sheet 1.

- *And finally, we have a function $\mathbb{P} : \mathcal{F} \to [0,1]$ satisfying $\mathbb{P}(\Omega) = 1$ and additivity for disjoint sets: if $A_1, A_2, \cdots \in \mathcal{F}$ are pairwise disjoint,*

$$\mathbb{P}(\bigcup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbb{P}(A_n).$$

This function $\mathbb{P}$ is called the probability

Notice the only differences are 1) we do not assume Ω to be finite 2) we assume that the set of events is stable under countable unions 3) we assume also the additivity of the probability under countable unions.

Exercise 1.1. *Show that each elementary probability space is a probability space.*

In fact probability spaces are an example of a general notion of measure spaces - probability spaces are just measure spaces with total mass equal to 1.

Definition 1.4 (Measure space, Borel 1898, Lebesgue 1901-1903)**.** *A measure space is a triple $(\Omega, \mathcal{F}, \mu)$, where*

- *Ω is a set, called the sample space or the universe.*
- *$\mathcal{F}$ is a set of subsets of Ω, satisfying:*
  - *$\emptyset \in \mathcal{F}$;*
  - *if $A \in \mathcal{F}$, then also $A^c \in \mathcal{F}$;*
  - *If $A_1, A_2, \cdots \in \mathcal{F}$, then also $\bigcup_{n\geq 1} A_n \in \mathcal{F}$.*

  $\mathcal{F}$ is called a σ-algebra and any $A \in \mathcal{F}$ is called a measurable set.
- *And finally, we have a function $\mu : \mathcal{F} \to [0, \infty]$ satisfying $\mu(\emptyset) = 0$ and countable additivity for disjoint sets: if $A_1, A_2, \cdots \in \mathcal{F}$ are pairwise disjoint,*

$$\mu(\bigcup_{n\geq 1} A_n) = \sum_{n\geq 1} \mu(A_n).$$

This function μ is called a measure. If $\mu(\Omega) < \infty$, we call μ a finite measure.

Geometrically we interpret:

- Ω as our space of points
- $\mathcal{F}$ as the collection of subsets for which our notion of volume can be defined
- μ our notion of volume: it gives each measurable set its volume.

It is important to make this link to measure theory as many properties of probability spaces directly come from there. Yet it is also good to keep in mind that probability theory is not just measure theory - as M. Kac has put it well, 'Probability is measure theory with a soul' and we adhere to this philosophical remark.

Remark 1.5. *You should compare the definition of a probability space / measure space with the definition of a topological space: there also we use a collection of subsets with certain properties to attach structure to the set. A question you should ask is: why do we use exactly countable unions and intersections for the events, and not finite or arbitrary?*

1.3 Some basic properties of probability spaces

We start by a few small remarks about the definition of a probability space:

Remark 1.6. *It is worth considering why ask for countable stability of the σ-algebra or countable additivity of the probability measure. Whereas this is more a meta-mathematical question, it is good to keep it in mind throughout the course. Let us here just offer two simple observations.*

First, countable sums naturally come up when we take limits of finite sums. In fact, countable additivity can be seen to be equivalent to certain form of continuity for the probability measure (see below).

Second, allowing for arbitrary unions leads easily to power-sets, and sums of uncountably many positive terms cannot be finite (see the exercise sheet).

Exercise 1.2. *Show that the countable additivity in the axioms of a probability space can be replaced with finite additivity plus the following statement: for any decreasing sequence of events $E_1 \supseteq E_2 \supseteq E_3 \dots$ we have that $\mathbb{P}(\cap_{i=1}^n E_i) \to 0$ as $n \to \infty$.*
$\star$ Does this hold in a general measure space?

Also we would like to remark another setting that explains well the usefulness of σ-algebras:

Remark 1.7. *Often in real life we only obtain information about the world step by step, and thus if we want to keep on working on the same probability space (which is helpful as then $\mathbb{P}$ will only need to be extended not redefined), we can consider a sequence of σ-algebras $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \dots$ called a filtration - each day we can ask some more yes/no questions because we already for example know what happened on the previous day and maybe also have learned something new. All possible information is contained in the power set $\mathcal{P}(\Omega)$.*

Probability spaces are usually classified in two types:

Definition 1.8 (Discrete and continuous probability spaces)**.** *Probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ with a countable sample space Ω are called discrete probability spaces and those with an uncountable sample space are called continuous probability spaces.*

In this course we will mainly work with discrete probability spaces, as they are technically easier to deal with. However, continuous probability spaces come up naturally and we won't be able to fully avoid them either.

Their technical difference can be summoned in the following proposition, whose non-examinable proof will be left for enthusiasts.

Proposition 1.9. *Let Ω be countable and $\mathcal{F}$ a $\sigma-$algebra on Ω. Then one can find disjoint events $E_1, E_2, \dots \in \mathcal{F}$ such that for every $E \in \mathcal{F}$ we can express $E = \cup_{i \in I_E} E_i$.*

Essentially, this says that for every discrete probability space it suffices to determine $\mathbb{P}(E_i)$ for a countable collection of disjoint sets E_i, and thereafter for every other set E we can use countable additivity to extend $\mathbb{P}$. Notice that this means it is first easy to check whether a given $\mathbb{P}$ satisfies all the axioms and even more importantly it is easy to check when two probability measures are equal.

For continuous probability spaces this does not necessarily hold - the useful σ-algebras are usually more complicated. To examplify why one doesn't want to necessarily use the power-set consider the following proposition, whose proof is in the appendix and relies on the axiom of choice:

Proposition 1.10. *There is no probability measure $\mathbb{P}$ on $([0,1], \mathcal{P}([0,1]))$ that is invariant under shifts, i.e. such that for any $A \in \mathcal{P}([0,1]), \alpha \in [0,1)$, we have that $\mathbb{P}(A + \alpha \mod 1) =$*

$\mathbb{P}(A)$, *where here we denote* $A + \alpha \mod 1 := \{a + \alpha \mod 1 : a \in A\}$, *the set obtained by shifting* A *by* α, *modulo* 1.

In fact, it comes out that the only way to remedy this situation is to make the relevant $\sigma-$algebra smaller. We would still want to be able to answer yes or no to questions like: is my random number equal to $\{x\}$ or is it in an interval (a, b)? Thanks to the fact that we have only countable additivity, this does not imply that our σ-algebra would need to be the power-set. And thanks to the properties of the $\sigma-$algebras, we can always construct at least some $\sigma-$algebra containing all our favourite sets - see the exercise sheet.

Let us now state some immediate consequences of the definitions about the $\sigma-$algebras and the probability measures:

Lemma 1.11 (Stability of the $\sigma - algebra$)**.** *Consider a set* Ω *with a* σ*-algebra* $\mathcal{F}$.

(1) If $A_1, A_2, \ldots, \in \mathcal{F}$, *then also* $\bigcap_{n\geq 1} A_n \in \mathcal{F}$.
(2) Then also $\Omega \in \mathcal{F}$ *and if* $A, B \in \mathcal{F}$, *then also* $A \setminus B \in \mathcal{F}$.
(3) For any $n \geq 1$, *if* $A_1, \ldots, A_n \in \mathcal{F}$, *then also* $A_1 \cup \cdots \cup A_n \in \mathcal{F}$ *and* $A_1 \cap \cdots \cap A_n \in \mathcal{F}$.

Proof of Lemma 1.11. By de Morgan's laws for any sets $(A_i)_{i\in I}$, we have that

$$\bigcap_{i\in I} A_i = (\bigcup_{i\in I} A_i^c)^c.$$

Property (1) follows from this, as if $A_1, A_2, \cdots \in \mathcal{F}$, then by the definition of a σ-algebra also $A_1^c, A_2^c, \cdots \in \mathcal{F}$ and hence

$$(\bigcup_{i\geq 1} A_i^c)^c \in \mathcal{F}.$$

For (3), again by de Morgan laws, it suffices to show that $A_1 \cup \cdots \cup A_n \in \mathcal{F}$. But this follows from the definition of a σ-algebra, as $A_1 \cup \cdots \cup A_n = \bigcup_{i\geq 1} A_i$ with $A_k = \emptyset$ for $k \geq n + 1$.

Point (2) is left as an exercise. □

In a similar vein, the basic conditions on the measure give rise to several natural properties:

Proposition 1.12 (Basic properties of a probability measure)**.** *Consider a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *Let* $A_1, A_2, \cdots \in \mathcal{F}$. *Then*

(1) For any $A \in \mathcal{F}$, *we have that* $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
(2) For any $n \geq 1$, *and* $A_1, \ldots, A_n$ *disjoint, we have finite additivity*

$$\mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) = \mathbb{P}(A_1 \cup \cdots \cup A_n).$$

In particular if $A_1 \subseteq A_2$ *then* $\mathbb{P}(A_1) \leq \mathbb{P}(A_2)$.
(3) If for all $n \geq 1$, *we have* $A_n \subseteq A_{n+1}$, *then as* $n \to \infty$, *it holds that* $\mathbb{P}(A_n) \to \mathbb{P}(\bigcup_{k\geq 1} A_k)$.
(4) We have countable subadditivity (also called the union bound): $\mathbb{P}(\bigcup_{n\geq 1} A_n) \leq \sum_{n\geq 1} \mathbb{P}(A_n)$.
(5) If for all $n \geq 1$, *we have* $A_n \supseteq A_{n+1}$, *then as* $n \to \infty$, *it holds that* $\mathbb{P}(A_n) \to \mathbb{P}(\bigcap_{k\geq 1} A_k)$.

Proof. Properties 1, 4 and second part of 2 were included in the Exercise sheet 1. The first part of property 2 follows like in the lemma above by taking $A_{n+1} = A_{n+2} = \cdots = \emptyset$ and using countable additivity.

So let us prove property 3: Write $B_1 = A_1$ and for $n \geq 2$, $B_n = A_n / A_{n-1}$. Then B_n are disjoint, $\bigcup_{n=1}^{N} B_n = A_N$ and $\bigcup_{n\geq 1} B_n = \bigcup_{n \geq 1} A_n$.

Thus by countable additivity

$$\mathbb{P}(\bigcup_{i\geq 1} A_i) = \mathbb{P}(\bigcup_{i \geq 1} B_i) = \sum_{i \geq 1} \mathbb{P}(B_i)$$

But $\mathbb{P}$ is non-negative, so

$$\sum_{i \geq 1} \mathbb{P}(B_i) = \lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{P}(B_i)$$

By countable additivity again

$$\sum_{i=1}^{n} \mathbb{P}(B_i) = \mathbb{P}(\bigcup_{i=1}^{n} B_n) = \mathbb{P}(A_n)$$

and (2) follows.

□

1.4 Random variables

In fact when studying a random phenomena we certainly don't want to restrict ourselves to yes and no questions. For example, in our model of a random number among $\{1, 2, \ldots, 12\}$ the natural question is not 'Is this number equal to 5?' but rather 'What number is it?'. Similarly in our example of discussing the weather, it is more natural to ask 'What is the temperature?', 'How much rain will there be in the afternoon?'?

Such numerical observations about our random phenomena will be formalised under the name of random variables. In essence they give a number for each state and thus as such are just functions $X : \Omega \to \mathbb{R}$ from the state-space to real numbers. However, we may not want to include all such functions for consistency reasons. Indeed, we want to be able to ask yes / no questions about our random numbers, e.g. Is the random number equal to 3? Is the temperature more than 18? But again the answer yes / no corresponds to certain subsets of states in the universe and as such should be events in our model. Thus there is a link between the collection of events, and and the collection of functions that can act as random variables. Let us without further give the general definition:

Definition 1.13 (Random variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We call a function $X : \Omega \to \mathbb{R}$ a random variable if for every interval (a, b) the set $X^{-1}((a, b)) := \{\omega \in \Omega : X(\omega) \in (a, b)\}$ is an event on the original probabiliuty space, i.e. belongs to $\mathcal{F}$.*

There is a simplification in the case of discrete probability spaces:

Lemma 1.14 (Random variables on discrete probability spaces)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a discrete probability space. Then $X : \Omega \to \mathbb{R}$ is a random variable if and only if for every $y \in \mathbb{R}$ we have that $X^{-1}(\{y\}) \in \mathcal{F}$.*

Proof. This can be verified carefully from the definitions and will be on the exercise sheet. □

For the structurally minded the definition of a random variable might look somewhat arbitrary. And indeed, I have been hiding one piece of information - the natural collection of events on $\mathbb{R}$ that we alluded to a little bit already in the previous subsection. We will directly state it on $\mathbb{R}^n$.

Definition 1.15 (Borel σ-algebra)**.** *The smallest σ-algebra on $\mathbb{R}^n$ that contains all open boxes of the form $(a_1, b_1) \times \cdots \times (a_n, b_n)$ is called the Borel σ-algebra. We denote it by $\mathcal{F}_B$*

Remark 1.16. *In fact this definition is even more general: given any topological space (X, τ), the smallest σ-algebra containing all open sets is called the Borel σ-algebra. You will see on the exercise sheet that this more general definition reduces to the previous one in the case of $\mathbb{R}^n$ with its Euclidean topology.*

Based on this an equivalent, possibly more structural definition of a random variable is as follows: a function $X : \Omega \to \mathbb{R}$ is a random variable if the preimage of every set in the Borel $\sigma-$algebra under X is an event. [5]

An important notion that comes with random variables is its law:

Lemma 1.17 (The law of a random variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable.*

Then there is a probability measure $\mathbb{P}_X$ induced on $(\mathbb{R}, \mathcal{F}_B)$ by defining $\mathbb{P}_X(F) := \mathbb{P}(X^{-1}(F)$ for every $F \in \mathcal{F}_B$. This probability measure $\mathbb{P}_X$ is called the law (or distribution) of a random variable X.

This is a lemma and not a definition as it needs to be proved that indeed $\mathbb{P}_X$ is a probability measure on $(\mathbb{R}, \mathcal{F}_B)$.

Proof of Lemma. We need to verify the axioms on a probability measure for a probability space:

- We have $\mathbb{P}_X(\mathbb{R}) = \mathbb{P}(\Omega) = 1$
- Similarly $\mathbb{P}_X(F) = \mathbb{P}(X^{-1}(F)) \in [0, 1]$ for all $F \in \mathcal{F}_B$
- Finally it remains to check countable additivity: let $F_1, F_2, \ldots$ be disjoint sets in $\mathcal{F}_B$. Then

  $$\mathbb{P}_X(\bigcup_{i\geq 1} F_i) = \mathbb{P}(X^{-1}(\bigcup_{i\geq 1} F_i)) = \mathbb{P}(\bigcup_{i\geq 1} X^{-1}(F_i)) = \sum_{i\geq 1} \mathbb{P}(X^{-1}(F_i)) = \sum_{i\geq 1} \mathbb{P}_X(F_i).$$

  Here we used the definition in the first and last equality, the properties of preimages in the second equality and the fact that $X^{-1}(F_i)$ are disjoint together with countable additivity in the third equality.

□

In words we showed that each random variable X induces a probability measure on the real numbers by just forgetting about the whole context and just concentrating on the number we see. For example in the case of weather in Lausanne, the temperature will give us a random variable and by just looking at its value and nothing else we have just a random real-valued number. Or more simply, if if we throw two fair coins and count the nunmber of heads, their sum will be a random variable that takes values in the set $\{0, 1, 2\}$. Thus the notion of the law of random variable gives us a way to compare random quantities arising in very different contexts.

Definition 1.18 (Equality in law)**.** *Let X, Y be two random variables defined possibly on different probability spaces. We say that X and Y are equal in law or equal in distribution, denoted $X \sim Y$ if for every $E \in \mathcal{F}_E$ we have that $\mathbb{P}_X(E) = \mathbb{P}_Y(E)$.*

[5] In measure theory such functions would be called measurable functions from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{F}_B)$; notice the similarity with the definition of continuous functions in your topology course.

We stress that when looking at the law of random variable the context gets forgotten - we only concentrate on the numerical value and the initial probability space $(\Omega, \mathcal{F}, \mathbb{P})$ only helps to determine $\mathbb{P}_X$ but plays no role thereafter. This means that we can nicely connect different random phenomena between each other. For example the indicator functions of all events that have probability p, independently on which probability space they have been defined, have the same law. Or more concretely, for example the following random variables have the same law:

- Number of heads in two independent tosses
- Number of prime factors when we choose uniformly a number among $\{1, 2, 3, 4\}$.

In some sense a large part of this course will be about studying and describing probability laws of random variables.

Section 2

Conditional probability and independence

In general, if we learn something new about our random phenomena, this knowledge influences and often changes our predictions for the rest of the model.

- For example in the case of a uniform random number between 1 and 12, if someone tells you that this number is even, then the probability of seeing 1 will suddenly be 0, but the probability of seeing 2 will rise from 1/12 to 1/6.
- In the case of weather in Lausanne, if someone tells us that it rains the whole day, then it is less likely to also be above 35 degrees.

The aim of this section is to set up the vocabulary to talk about how the knowledge about some event or random variable influences the probabilities we should assign to other events. This leads us to talk about conditional probabilities and to discuss the case where events do not influence each other, giving rise to an important notion of probability theory called independence.

2.1 Conditional probability

We have already considered (in the course and on the example sheets) many unpredictable situations where several events naturally occur either at the same time or consecutively: a sequence of coin tosses or successive steps in a random walk, or different links or edges in a random graph. In all these cases, the fact that one event has happened could easily influence the others. For example, if you want to model the financial markets tomorrow, it seems rather advisable to take into account what happened today. To talk about the change of probabilities when we have observed something, we introduce the notion of conditional probability:

Definition 2.1 (Conditional probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $E \in \mathcal{F}$ with $\mathbb{P}(E) > 0$. Then for any $F \in \mathcal{F}$, we define the conditional probability of the event F given E (i.e. given that the event E happens), by*

$$\mathbb{P}(F|E) := \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(E)}.$$

Recall that $E \cap F$ is the event that both E and F happen. Hence, as the denominator is always given by $\mathbb{P}(E)$, the conditional probability given E is proportional to $\mathbb{P}(E \cap F)$ for any event F. Here is the justification for dividing by $\mathbb{P}(E)$:

Lemma 2.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $E \in \mathcal{F}$ with $\mathbb{P}(E) > 0$. Then $P(\cdot|E)$ defines a probability measure on $(\Omega, \mathcal{F})$, called the conditional probability measure given E.*

Proof. First, notice that $\mathbb{P}$ is indeed defined for every $F \in \mathcal{F}$. Next, $\mathbb{P}(\emptyset|E) = \mathbb{P}(\emptyset)/\mathbb{P}(E) = 0$ and $\mathbb{P}(\Omega|E) = \mathbb{P}(E)/\mathbb{P}(E) = 1$. So it remains to check countable additivity.

So let $F_1, F_2, \ldots \mathcal{F}$ be disjoint. Then also $E \cap F_1, E \cap F_2, \ldots$ are also disjoint. Hence

$$\mathbb{P}(\bigcup_{i\geq 1} F_i|E) = \frac{\mathbb{P}((\bigcup_{i\geq 1} F_i) \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(\bigcup_{i\geq 1}(F_i \cap E))}{\mathbb{P}(E)} = \sum_{i\geq 1} \frac{\mathbb{P}(F_i \cap E)}{\mathbb{P}(E)} = \sum_{i\geq 1} \mathbb{P}(F_1|E),$$

and countable additivity follows.

□

It should be remarked that conditional probability of an event might sometimes be similar to the initial probability (we will see more about this very soon), but it might also be drastically different. A somewhat silly but instructive example is the following:

- Conditional probability of the event E^c, conditioned on E is always zero, no matter what the original probability was;
- similarly the conditional probability of E, conditioned on E is always 1.

Or for a more senseful exercise consider the following:

Exercise 2.1 (Random walk and conditional probabilities)**.** *Consider the simple random walk of length n.*

- *What is the probability that the walk ends up at the point n at time n? Now, suppose that the first step was -1. What is the probability that the walk ends up at the point n at time n now?*
- *Suppose that n is even. What is the probability that the walk ends up at the point 0 at time n? Now, suppose that the first step was -1. What is the probability that the walk ends up at the point 0 at time n now?*

One also has to be very careful about the exact conditioning, as two similarly sounding conditionings can induce very different conditional probabilities. In general, we need to know something extra about the relation of two events to know how the probability of one changes when conditioned on the other.

There are some cases where these relations and thus conditional probabilities are easy:

- When $E \subseteq F$, then the conditional probability of F given E is just 1.
- When $F \subseteq E^c$, then the conditional probability of F given E is just 0.
- The third case is when F and E are so called independent: in that case $\mathbb{P}(F|E) = \mathbb{P}(E)$ basically by definition (we will come back to that).

In general, there are not many tools to calculate conditional probabilities, but there is one very useful tool called the Bayes' formula or the Bayes' rule:

2.1.1 Bayes' rule

Proposition 2.3 (Bayes' rule)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and E, F two events of positive probability. Then*

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(F|E)\mathbb{P}(E)}{\mathbb{P}(F)}$$

We will discuss this at a greater length next week.