Basics on probability and statistics

Meryem Benammar

18 octobre 2021

Résumé

Probability theory is made to predict, or bring a prior knowledge, on an experiment before observing the output of the experiment. In the present course book, we will list a few basic elements of probability theory, with a specific emphasis on the practical implications of the notions and results obtained. This book is not meant to be a comprehensive textbook on probabilities and statistics, but is rather a reminder of some definitions, and an introduction to the different notations used later in information engineering.

Table des matières

1 From probability spaces to random variables

In the following we define and list a few basic elements on events, probability spaces and probability measures.

1.1 Experiments, events and sets

Consider an experiment of which we have not yet observed the output, and let us characterize this experiments by means of events. To this end, assume that the output of all experiments are elements of an alphabet Ω. For instance, consider each of the following experiments :

- Tossing a coin : $\Omega = \{\text{Head, Tail}\}$
- Throwing a dice, then $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Throwing two dice, then $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$

Once that we have introduced the alphabets in which the experiments take their values, we need to define the notion of event, which we will denote E. For instance, in a dice throw, we could define many events :

- Obtain the value 1, denoted by $E_1 = \{1\}$
- Obtain an even value, denoted by $E_2 = \{2, 4, 6\}$
- Obtain an odd value, denoted by $E_3 = \{1, 3, 5\}$
- Obtain a value grater than or equal to 3, which we will denote $E_4 = \{3, 4, 5, 6\}$
- Obtain a value smaller strictly than 3 $E_5 = \{1, 2\}$

An event can be understood as the set of values we can obtain from an experiment.

As can be seen, an event E is in fact a subset of the alphabet Ω, and thus, one can imagine all sorts of combination of events :

- Obtain an even value, greater than 4, which implies $E_6 = \{2, 4, 6\} \cap \{4, 6\} = \{4, 6\}$
- Obtain a value grater than or equal to 3, and strictly smaller than 5 which yields $E_7 = \{3, 4, 5, 6\} \cap \{3, 4\} = \{3, 4\}$
- Obtain a value smaller strictly than 3 or odd $E_8 = \{1, 2\} \cup \{1, 3, 5\} = \{1, 2, 3, 5\}$

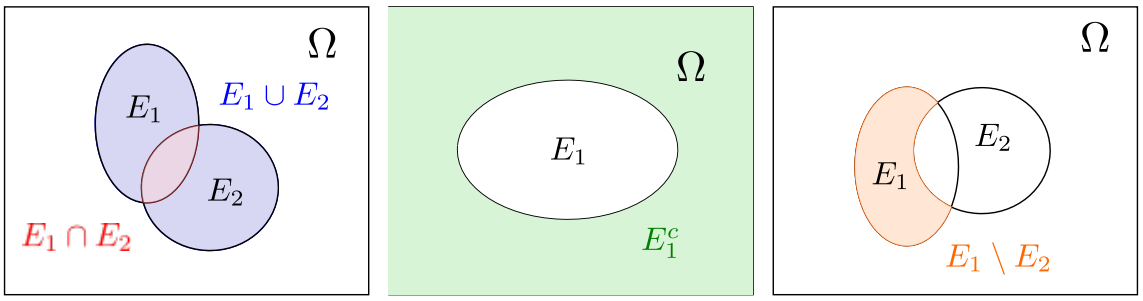

FIGURE 1 – Set and event operations

Thus, all operations on events can be thought as operations on sets (set axioms), as shown in Table 1.

Operation on event	Operation on sets
Event 1 and Event 2	$E_1 \cap E_2$
Event 1 or Event 2	$E_1 \cup E_2$
Not event 2	$E_2^c = \Omega \setminus E_2$
Event 1 except Event 2	$E_1 \setminus E_2$
Null event	$\emptyset$
Trivial event	Ω

TABLE 1 – Set axioms

In the following, we will denote the set of all possible events $\mathcal{E}$ as the *event space*, which consists then in the set of all possible subsets of the set Ω.

Exercise 1 *: NASA's Deep Space Network (DSN) consists in three antennas deployed in three locations : Madrid, Goldstone and Canberra. Consider a rover which wants to initialize a first communication link with the ground through the DSN by connecting at random to one or more of the antennas. Describe the alphabet Ω and event space $\mathcal{E}$ (list all possible events).*
In order for the communication link to be reliable, the rover needs to connect to at least two antennas of the DSN. Descrive this event E (in terms of possible elements).

1.2 A probability space

Now that we have defined an alphabet, an event, an event space, let us define a probability measure as follows.

Definition 1 (Probability measure)

Let Ω be an alphabet, and let $\mathcal{E}$ be a event space of this alphabet (all possible subsets of it). Then, we can define a probability measure $\mathbb{P}$ as a function defined on the event set

$$\begin{aligned} \mathbb{P} : \mathcal{E} &\rightarrow [0,1] \\ E &\rightarrow \mathbb{P}(E) \end{aligned}$$

where $\mathbb{P}$ verifies the following axioms

— *Positivity : for all events E, their probability is positive, i.e.,*

$$\mathbb{P}(E) \geq 0, \quad \text{for all } E \in \mathcal{E} \tag{1}$$

— *Exhaustivity : the probability of the trivial event is 1, i.e.,*

$$\mathbb{P}(\Omega) = 1 \tag{2}$$

— *Additivity : if two events E_1 and E_2 are disjoint, their probabilities add up, i.e.,*

$$E_1 \cap E_2 = \emptyset \quad \Rightarrow \quad \mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2). \tag{3}$$

Properties 1 *Let Ω be an alphabet, and let $\mathcal{E}$ be a event space of this alphabet, and $\mathbb{P}$ a probability measure defined on $\mathcal{E}$. Given the three axioms of probability theory, we can prove that :*

— *For all event $E \in \mathcal{E}$,*

$$\mathbb{P}(E^c) = 1 - \mathbb{P}(E) \tag{4}$$

— *For all pairs of events (E_1, E_2) in $\mathcal{E}$,*

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2) \tag{5}$$

— *If the elements of $\mathcal{X}$ are indexed by x and are countable, then*

$$\sum_{w \in \Omega} \mathbb{P}(w) = 1 \tag{6}$$

— *The probability of the null event is zero, i.e.,*

$$\mathbb{P}(\emptyset) = 0 \tag{7}$$

— *For all events E with countable elements,*

$$\mathbb{P}(E) = \sum_{w \in E} \mathbb{P}(w) \tag{8}$$

— *For any pair of events E_1 and E_2,*

$$E_1 \subset E_2 \quad \Rightarrow \quad \mathbb{P}(E_1) \leq \mathbb{P}(E_2) \tag{9}$$

Exercise 2 *: Based on the axioms of probability theory, prove the previous properties.*

Now that we have define a probability measure, let us define the notion of a probability space to describe fully an experiment, the outputs of interest, and the probability with which we can obtain these outputs.

Definition 2 (Probability space)

A probability space $\mathcal{P} = (\Omega, \mathcal{E}, \mathbb{P})$ is defined by three components

— *An alphabet set Ω representing a set of all possible values*

— *An event set $\mathcal{E}$ consists in all possible subsets of the alphabet Ω*

— *A probability measure $\mathbb{P}$*

1.3 Random variables as mappings

Now that we have defined a probability space, we can define a random variable and a random vector. In an experiment, the output observed might not be what is truly important to us, but rather a function of this output. For instance, while throwing a dice, what is important to us might not be the value of the dice itself, but whether it is an odd or even value. When throwing two dice, what is important to us might not be the value

of each dice, but rather their sum, or whether they are equal or not (like in Parchisi). Of course, the value of interest might be the very output of the experiment. A broad definition of a random variable is then a mapping from a probability space to another as shown in Figure 2.

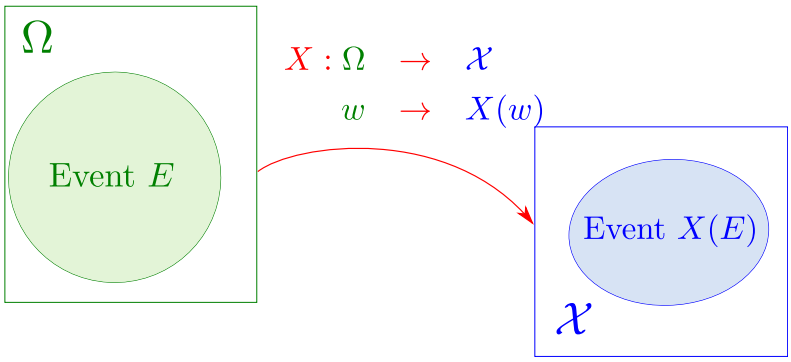

FIGURE 2 – Random variable definition

Definition 3 *Let us the a probability space $\mathcal{P} = (\Omega, \mathcal{E}, \mathbb{P})$ and $(\mathcal{X}, \mathcal{E}_X)$ a new sample set and its event space. The mapping from Ω to $\mathcal{X}$ is called a random variable which we will denote as X :*

$$X : \quad \Omega \rightarrow \mathcal{X} \tag{10}$$
$$w \rightarrow X(w). \tag{11}$$

To such a random variable X is associated a probability measure $\mathbb{P}_X$ defined, for all events E_X in $\mathcal{E}_X$

$$\mathbb{P}_X(E_x) = \mathbb{P}(X^{-1}(E_X)) = \mathbb{P}(\{w \in \Omega, X(w) \in E_x\} \tag{12}$$

The random variable is thus defined by the new probability space $\mathcal{P}_X = (\mathcal{X}, \mathcal{E}_X, \mathbb{P}_X)$.

The notion of random variable as introduced herebefore is very general, let us now give examples of possible random variables defined on simple experiments.

Exercise 3 *: Consider the experiment of throwing a dice, i.e., $\Omega = \{1, ..., 6\}$. Consider as well that the dice is fair, and so, all faces are equally probable. On this simple experiment, we can define a variety of random variables X. For each of the suggested random variable, write the corresponding probability space $\mathcal{P}_X = (\mathcal{X}, \mathcal{E}_X, \mathbb{P}_X)$ and characterize the probability measure $\mathbb{P}_X$.*

- *The parity of the obtained result, i.e., whether it is even or odd.*
- *Whether the result is greater or samller than 3*
- *The value of the result modulo 4 (the remainder of the division of the result by 4)*
- *The square of the result*
- *The value of the result itself*

In the following, when working with random variables, we will mostly rely on the probability space and probability measure of the random variable itself $\mathcal{P}_X = (\mathcal{X}, \mathcal{E}_X, \mathbb{P}_X)$, and not the underlying experiment.

However, from a notation point of view, we will keep the dependence on the experiment through the notion of *realizations*, which is basically $X(w)$. Hence, throughout this text book, random variables are always denoted by capital letters X while their realizations are denotes by regular case letters $x = X(w)$. Hence, we say that a realization of the random variable X is x, if the result of the experiment yields $X(w) = x$.

2 Discrete random variables

Let X be a random variable defined over the sample set $\mathcal{X}$ (set of all possible realizations of X) and with associated probability measure $\mathbb{P}_X$. Depending on whether the support set is continuous (e.g. $\mathcal{X} = \mathbb{R}$), or discrete (e.g. $\mathcal{X} = [1 : M]$), we will distinguish between discrete and continuous random variables. In this section, we focus on *discrete* random variables.

2.1 Probability mass function (pmf)

Let X be a discrete random variable defined over the sample set $\mathcal{X}$ (set of all possible realizations of X is finite) and with associated probability measure $\mathbb{P}_X$.

Definition 4 (Probability mass function (pmf)) *The probability mass function of a discrete random variable X with associated probability measure $\mathbb{P}_X$ is defined by*

$$\begin{aligned} P_X : \quad \mathcal{X} &\rightarrow [0:1] && (13)\\ x &\rightarrow P_X(x) = \mathbb{P}_X(X = x) = \mathbb{P}_X(\{x\}) && (14) \end{aligned}$$

Unlike the probability measure $\mathbb{P}_X$ which can be defined on all possible events of $\mathcal{X}$, i.e., all possible subsets E_X of $\mathcal{X}$, the pmf is defined only on singletons $\{x\}$, and is yet sufficient to characterize the probabilities of all events of $\mathcal{X}$.

We have by definition of the pmf that

$$\sum_{x \in \mathcal{X}} P_X(x) = 1. \tag{15}$$

Examples 1 *Characterizing the pmf of a random variable might not be an easy task, but often, the random variables we ecounter in the physical world, are distributed following some known pmfs. Here is a list of such known discrete pmf.*

1. *Bernoulli of parameter p*
2. *Discrete uniform over the interval $[1 : K]$*
3. *Binomial with parameter p*
4. *Constant variable equal to K*

Exercise 4 *: For each of the pmf cite above, characterize :*

- *The sample set* $\mathcal{X}$
- *The corresponding pmf* P_X
- *Plot the pmf* P_X *(using bars if necessary)*
- *An experiment in which you encouter this random variable*

2.2 Moments : expectation and variance

When describing a random variable, the pmf P_X alone captures all the necessary information of the random variable. However, this pmf is sometimes too informative, sincce it gives the values $P_X(x)$ for all $x \in \mathcal{X}$, and thus, depending on the number of elements in $\mathcal{X}$, this pmf might be hard to analyse. Sometimes, a simple numerical value extracted from this pmf is satisfactory for what we need, alike the average of the random variable, or the standard deviation from the average. These two simple numerical values are called *1st order* and *2nd order* moments of the random variable.

Definition 5 (Moments) *To each random variable* X *with pmf* P_X *are associated*

— *An expected (average) value* $\mathbb{E}(X)$ *(first order moment)*

$$\mathbb{E}(X) \triangleq \sum_{x \in \mathcal{X}} x.P_X(x) \tag{16}$$

— *A variance (squared standard deviation)* $\mathbb{V}(X)$ *(second order moment)*

$$\mathbb{V}(X) \triangleq \mathbb{E}\left(X^2\right) - \mathbb{E}^2(X). \tag{17}$$

The moments of a random variable have the following properties.

Properties 2 (Moments)

1. *The expectation is linear, i.e.,*

$$\mathbb{E}(f(X)) = f\left(\mathbb{E}(X)\right) \tag{18}$$

for any linear transformation f*.*

2. *For any constant* α*, we have that*

$$\mathbb{V}(\alpha.X) = \alpha^2 \mathbb{V}(X).$$

3. *For any constant* c*, we have that*

$$\mathbb{V}(X + c) = \mathbb{V}(X).$$

Exercise 5 *: Compute the expectation and variance of each of the previous classical pmfs.*

3 Discrete random vectors

Let us recall the experiment in which we threw two dice at once. The results of this experiment consist in the all possible pairs of values in $\{1, ..., 6\} \times \{1, ..., 6\}$, and can hence be considered as *a pair* of random variables which we will call X and Y. Alike the case in which we had a single random variable X, we can define a probability measure for the pair (X, Y), which we will denote as $\mathbb{P}_{X,Y}$ and name *joint* pmf. Now assume that in this experiment, what is interesting for us is whether the output of the first dice is faire. In this case, we can extract from the joint events (x, y) only the observation of the events x, by discarding the result of the second dice, and hence end up with a pmf on the first dice throw P_X. We could proceed similarly for the second dice. In this case, we say that we have extracted the marginal pmfs P_X and P_Y from the joint pmf $P_{X,Y}$.

In the following, we introduce the concepts herebefore introduced, more formally.

3.1 Joint and marginal pmfs

Let (X, Y) be a pair of random variables defined over the product support sample set $\mathcal{X} \times \mathcal{Y}$, with an iduced probability measure $\mathbb{P}_{X,Y}$.

Definition 6 (Joint and marginal pmfs)

The joint pmf $P_{X,Y}$ associated with the pair (X, Y) is given by

$$\begin{array}{rcl} \mathcal{X} \times \mathcal{Y} & \rightarrow & [0, 1] \\ (x, y) & \rightarrow & P_{X,Y}(x, y) = \mathbb{P}_{X,Y}(X = x \text{ and } Y = y) \end{array}$$

To the joint pmf $P_{X,Y}$ are associated two marginal pmfs P_X and P_Y defined by

$$P_X(x) = \sum_{y \in \mathcal{Y}} P_{X,Y}(x, y) \ , \ P_Y(y) = \sum_{x \in \mathcal{X}} P_{X,Y}(x, y) \tag{19}$$

A joint pmf $P_{X,Y}$ can be thought of as a matrix with $|\mathcal{X}|$ (if finite) lines and $|\mathcal{Y}|$ (if finite) columns, which associates to each pair of possible realizations (x, y), the value $P_{X,Y}(x, y)$.

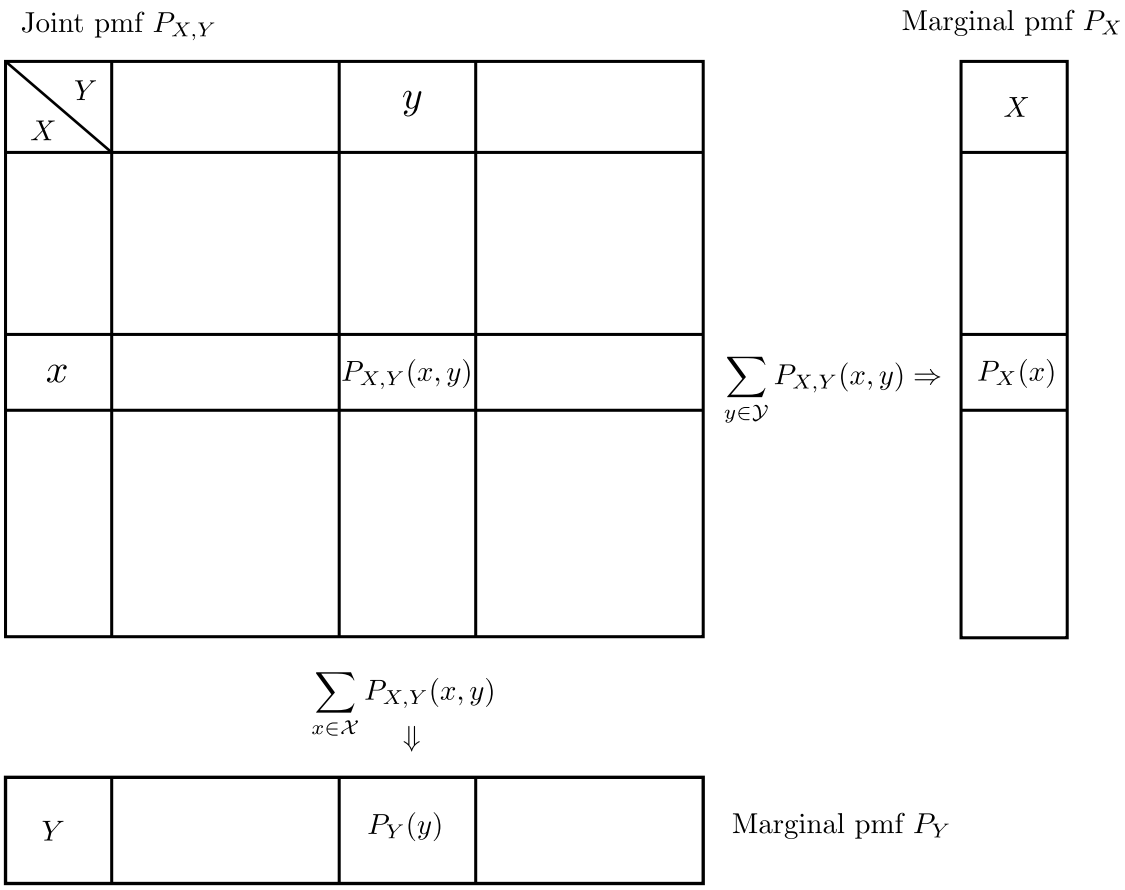

FIGURE 3 – Joint and marginal pmf

Properties 3

The joint and marginal distributions verify the following properties.

1. *Exhaustivity of joint pmfs*

$$\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y) = 1. \tag{20}$$

2. *Exhaustivity of marginal pmfs*

$$\sum_{x\in\mathcal{X}} P_X(x) = 1 \text{ and } \sum_{y\in\mathcal{Y}} P_Y(y) = 1 \tag{21}$$

3. *Independence : X and Y are independent random variables iif,*

$$\forall(x,y) \qquad P_{X,Y}(x,y) = P_X(x)P_Y(y) \tag{22}$$

3.2 Conditional pmfs

Let us recall now the experiment of throwing two dice, which we described by a pair of random variables (X,Y). Assume now that we are interested in characterizing the probability of obtaining all values x for the first dice, knowing that we have aleady observed a signle and given value y of the second dice. The obtained pmf on all possible values x knowing a value y, is called a conditional pmf and is denoted by $P_{X|Y}(\cdot|y)$.

Definition 7 (Conditional pmfs)
The conditional pmfs associated with $P_{X,Y}$ are defined by

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} = \mathbb{P}(X = x|Y = y)$$
$$P_{Y|X}(y|x) = \frac{P_{X,Y}(x,y)}{P_X(x)} = \mathbb{P}(Y = y|X = x)$$

Conditional pmf $P_{X|Y}$

X \ Y		y	
x		$P_{X\|Y}(x\|y)$	

$$\Downarrow$$
$$\sum_{x\in\mathcal{X}} P_{X|Y}(x|y) = 1$$

FIGURE 4 – Conditional pmf

Properties 4

The conditional pmfs verify the following properties.

1. *Independence : X and Y are independent random variables iif,*

$$\forall(x,y) \qquad P_{X|Y}(x|y) = P_X(x) \quad \text{and} \quad P_{Y|X}(y|x) = P_Y(y) \tag{23}$$

2. *Exhaustivity of conditional pmf*

$$\sum_{x\in\mathcal{X}\times\mathcal{Y}} P_{X|Y}(x|y) = 1. \tag{24}$$

" The joint pmf rules them all " : if you have the joint law $P_{X,Y}$, then you can define the marginal pmfs P_X and P_Y, as well as the conditional pmfs $P_{X|Y}$ and $P_{Y|X}$. However, if you know only the marginals, or only the conditional pmfs, you can build as many joint pmf as you want.

Examples 2 (Classical channel models) *A communication channel is a random transformation which transforms the channel input X into a channel output Y with a conditional probability $P_{Y|X}$. In the following, we give two classical examples of channels encountered in communication systems, namely, the Binary Symmetric Channel (BSC) and the Binary Erasure Channel (BEC).*

The Binary Symmetric Channel $BSC(p)$ is defined by $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $P_{X,Y}$,

Y, X	0	1
0	$\frac{1-p}{2}$	$\frac{p}{2}$
1	$\frac{p}{2}$	$\frac{1-p}{2}$

$\Rightarrow$

$Y\|X$	0	1
0		
1		

p is called the crossover probability. This channel model can also be written in a additive form,

$$Y = X \oplus W \tag{25}$$

where X is a $Bern(0.5)$ random variable, W is a $Bern(p)$ random variable, X and W are independent, and $\oplus$ is the binary XOR operation.

Exercise 6 *: Given $P_{X,Y}$ of the $BSC(p)$ channel, compute $P_{Y|X}$ for all possible (x, y). Check that the additive channel model yields the same conditional probability.*

Assume that $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, E\}$ and that $P_{X,Y}$ is given in the table

Y, X	0	1
0	$\frac{1-e}{2}$	0
1	0	$\frac{1-e}{2}$
E	$\frac{e}{2}$	$\frac{e}{2}$

$\Rightarrow$

$Y\|X$	0	1
0		
1		
E		

e is called an erasure probability.

Exercise 7 *: Given $P_{X,Y}$ of the $BEC(e)$, compute $P_{Y|X}$ for all possible (x, y). Hint : compute first the marginal law P_X.*

3.3 Bayes' formula

Bayes' formula is very useful in everyday engineering probabilities. Indeed, often we have input data X whose pmf is known, we call it the *prior* distribution. We often know as well the process by which the observable Y is generated from the original data X which we denote as $P_{Y|X}$ and often call the conditional observation pmf or *transition pmf*. However, what is of interest for us, is having observed a certain y, what is the probability of the original data being a given x, i.e., what is interesting is the pmf $P_{X|Y}$. This pmf is call

the *posterior* probability. The relationship between the prior P_X, the transision pmf $P_{Y|X}$ and the posterior $P_{X|Y}$ is given by Bayes' formula, which writes as follows :

$$P_{X|Y}(x|y) = \frac{P_X(x)P_{Y|X}(y|x)}{\sum_{x'} P_X(x')P_{Y|X}(y|x')} \tag{26}$$

Exercise 8 : *Prove Bays' formula.*

3.4 Discrete random vectors

Let be $(X_1, \ldots, X_n)$ a vector of n random variables, defined over the support set $\mathcal{X}_1 \times \ldots \mathcal{X}_n$. Similarly to pairs of random variables, we can define a joint pmf and a collection of marginal and conditional pmfs.

3.4.1 Joint and marginal pmfs

Definition 8 (Joint and marginal pmfs)

The joint pmf of the vector can be defined as $P_{X_1,\ldots,X_n}$

$$\begin{array}{rcl} \mathcal{X}_1 \times \ldots \mathcal{X}_n & \to & [0,1] \\ (x_1, \ldots, x_n) & \to & P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \end{array}$$

To the joint pmf $P_{X_1,\ldots,X_n}$ *are associated* n *marginal pdfs*

$$P_{X_i}(x_i) = \sum_{(x_1,\ldots,x_n)\setminus x_i} P_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

In the following, we state one main property which is satisfied by the joint pmf of a vector, and which will be crucial in this course, namely, the chain rule.

3.4.2 The chain rule

Properties 5 (The chain rule)

The joint pmf can be expanded using the so-called chain rule as follows

$$\begin{aligned} P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= \prod_{i=1}^{n} P_{X_i|X_1,\ldots,X_{i-1}}(x_i|x_1, \ldots, x_{i-1}) \\ &= \prod_{i=1}^{n} \frac{P_{X_1,\ldots,X_i}(x_1, \ldots, x_i)}{P_{X_1,\ldots,X_{i-1}}(x_1, \ldots, x_{i-1})} \end{aligned}$$

Example 1 (iid variables) *A set of variables* $(X_1, \ldots, X_n)$ *are deemed pairwise independent if their joint pmf verifies*

$$P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} P_{X_i}(x_i),$$

If, further, the variables are identically distributed, i.e., they follow the same law P_X, then

$$P_{X_1,\dots,X_n}(x_1,\dots,x_n) \quad = \quad \prod_{i=1}^{n} P_X(x_i).$$

Such variables are said to be independent and identically distributed (iid). In the following, we will often encounter such vectors of iid variables.

4 Continuous random variables

Unlike discrete random variables where the sample set $\mathcal{X}$ was a discrete (countable) set, a continuous random variable has a sample set which is not countable, and we will always assume that is the set of real numbers $\mathbb{R}$.

Almost all definitions in the discrete random variable case have their equivalent for continuous random variables. We detail in the following the main tools used to characterize continous random variables (and vectors) and highlight, whenever relevant, the differences with discrete random variables.

4.1 Probability density function (pdf)

In the continuous case, the rigorous definition of a probability measure invokes more intricate results than the intuitive definition of the pmf. The equivalent measure to the pmf in the continuous case is called the *probability density function* (pdf) and is defined as follows.

Definition 9 (Probability density function)
Let X be a continous random variable ($X : \omega \to \mathbb{R}$), then the associated probability density function is denoted by f_X and verifies the following properties :

- *Positivity :*
$$\forall x \in \mathbb{R}, \quad f_X(x) \geq 0 \tag{27}$$
- *Exhaustivity :*
$$\int_{x\in\mathcal{X}} f_X(x)\ dx = 1 \tag{28}$$
- *Interval probability*
$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\ dx \tag{29}$$

The exhaustivity of the pdf implies that, unlike discrete variables, for continuous variables it is the area below the pdf f_X which allows to characterize the probability events of the continuous variable.

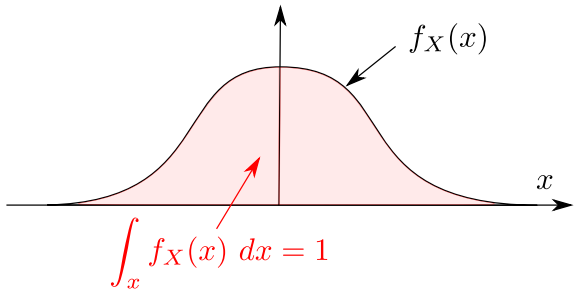

FIGURE 5 – Probability density function

For a continuous random variable, the probability of X being equal to a given value x is equal to 0. This can be seen easily from the following

$$\mathbb{P}(X = a) = \mathbb{P}(a \leq X \leq b) = \int_a^a f_X(x)\ dx = 0 \tag{30}$$

As such, unlike the pmf of a discrete random variable, the pdf $f_X(x)$ of a continuous random variable should never be understood as the probability that $X = x$.

4.2 Moments : expectation and variance

Definition 10 (Moments)

To each random variable X with pmf P_X/ pdf f_X are associated

— *An expected value $\mathbb{E}(X)$ (first order moment[a])*

$$\int_{x\in\mathcal{X}} x.f_X(x)\ dx. \tag{31}$$

— *A variance $\mathbb{V}(X)$ (second order moment)*

$$\mathbb{V}(X) \triangleq \mathbb{E}\left(X^2\right) - \mathbb{E}^2(X). \tag{32}$$

a. The notation $\triangleq$ is used when defining a notion for the first time.

Properties 6 (Moments) *The expectation and variance have the following properties*

1. *The expectation is a linear transformation, i.e.,*

$$\mathbb{E}(f(X)) = \int_x f_X(x) f\left(X\right)\ dx \tag{33}$$

for any linear transformation f.

2. *For any constant α, we have that*

$$\mathbb{V}(\alpha.X) = \alpha^2\mathbb{V}(X).$$

Examples 3 *In the following, we list some examples of probability laws which will be of interest throughout this course.*

1. *Gaussian with mean μ and variance σ^2*
2. *Exponential law with parameter λ*
3. *Continuous uniform over an interval $[a, b]$*

Exercise 9 *: For each of these laws,*

- *describe a physical process in which you can encounter the law*
- *give the formula of the pdf*
- *compute the expectation*
- *compute the variance.*

4.3 Cumulative distribution function (cdf)

Unlike discrete random variables, continuous random variables admit what we denote a cumulative distribution function (cdf). The cdf of a random variable can be defined as follows.

Definition 11 (Cumulative distribution function (cdf)) *Let X be a continuous random variable with associated pdf f_X. Then, to X is associated a cumulative distribution function (cdf) F_X defined by*

$$F_X(a) = \mathbb{P}(X \leq a) = \int_{-\infty}^{a} f_X(x)\, dx \tag{34}$$

The cdf $F_x(a)$ can be interpreted as the area below the pdf up to a certain value a as shown in Figure 6.

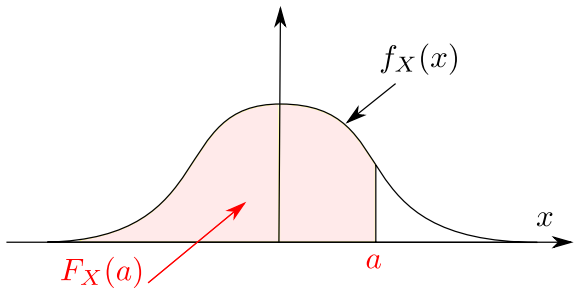

FIGURE 6 – Cumulative distribution function : integration

Properties 7 *The cdf F_X of a continuous random variable verifies the following*

1. *Exhaustivity*

$$F_X(\infty) = 1 \tag{35}$$

2. *Increasing function*

$$a_1 \leq a_2 \Rightarrow F_X(a_1) \leq F_X(a_2) \tag{36}$$

3. *Positivity*

$$F_X(-\infty) = 0 \tag{37}$$

4. Relation to pdf

$$f_X(x) = \frac{\partial F_X}{\partial x}(x) \tag{38}$$

An example of cdf, satisfying thus the aforementioned properties, is given in Figure 7.

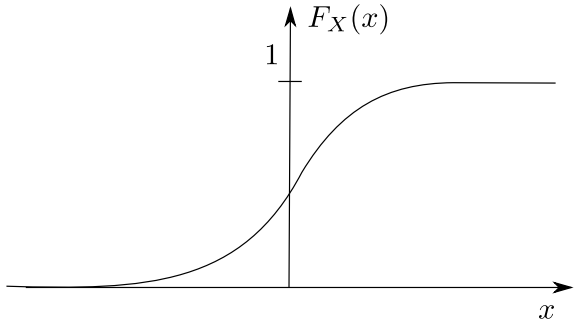

FIGURE 7 – Cumulative distribution function

4.4 Continuous random vectors

Similarly to discrete random vectors, we can define as well continous random vectors, with the same properties and behaviours as discrete random vectores.

Let be $(X_1, \ldots, X_n)$ a vector of n continuous random variables, defined over the support set $\mathcal{X}_1 \times \ldots \mathcal{X}_n$.

Definition 12 (Joint and marginal pds)

The joint pdf of the vector can be defined as $f_{X_1,\ldots,X_n}$

$$\begin{array}{rcl} \mathcal{X}_1 \times \ldots \mathcal{X}_n & \rightarrow & [0,1] \\ (x_1, \ldots, x_n) & \rightarrow & f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \end{array}$$

To the joint pdf $f_{X_1,\ldots,X_n}$ *are associated n marginal pdfs*

$$f_{X_i}(x_i) = \int_{(x_1,\ldots,x_n)\setminus x_i} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\ dx_1 ... dx_n$$

In the following, we state one main property which is satisfied by the joint pdf of a random vector, and which will be crucial in this course, namely, the chain rule.

Properties 8 (The chain rule)

The joint pdf can be expanded using the so-called chain rule as follows

$$\begin{array}{rcl} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) & = & \prod_{i=1}^{n} f_{X_i|X_1,\ldots,X_{i-1}}(x_i|x_1, \ldots, x_{i-1}) \\ & = & \prod_{i=1}^{n} \frac{f_{X_1,\ldots,X_n}(x_1, \ldots, x_i)}{f_{X_1,\ldots,X_{i-1}}(x_1, \ldots, x_{i-1})} \end{array}$$

Example 2 (iid variables) *A set of variables $(X_1, \dots, X_n)$ are deemed pairwise independent if their joint pdf verifies*

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i),$$

If, further, the variables are identically distributed, i.e., they follow the same law f_X, then

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} f_X(x_i).$$

Such variables are said to be independent and identically distributed (iid). In the following, we will often encounter such vectors of iid variables.

5 Basics on statistics

Statistics deal as well with random variables, but from another perspective. When probabilities aim at predicting the value of an experiment based on some knowledge on the underlying process, statistics aim at learning the underlying process based on observations of the experiment.

5.1 Law of large numbers

In the following, we state one of the main results of probability theory and statistics which will be used extensively in Shanon's information theoretic theorems.

Theorem 1 (Law of Large Numbers (LLN))

Let $(X_1, \dots, X_n)$ be n iid random variables, with pmf P_X/pdf f_X, and let $\mu = \mathbb{E}(X)$ be the expectation of X.

The empirical mean $\bar{X}_n$ of $(X_1, \dots, X_n)$, defined by

$$\bar{X}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i \tag{39}$$

converges in probability, as n, grow infinite, to μ, i.e.,

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu \right) = 1. \tag{40}$$

The LLN allows us, from iid samples of a given distribution, to estimate its mean μ. The empirical mean $\bar{X}_n$ is deemed a good estimator of the expectation of a draw of iid variables.

5.2 Central limit theorem

In the following, we describe the behavior of iid random variables as their number grows infinite, and relate to practical implementation of this results.

Theorem 2 (Central limit theorem (CLT))
Let $(X_1, \ldots, X_n)$ be n iid random variables, with pmf P_X/pdf f_X, and let $\mu = \mathbb{E}(X)$ be the expectation of X and σ^2 be its variance.

The random variable $Z_n = \sqrt{n}(\bar{X}_n - \mu)$ defined by

$$Z_n = \frac{\sqrt{n}}{n} \sum_{i=1}^{n} X_i - \mu \tag{41}$$

converges in distribution, as n, grow infinite, to a normal Gaussian distribution

$$\lim_{n \to \infty} \mathbb{P}\left(Z_n \leq z\right) = \Phi(z). \tag{42}$$

where $\Phi(z)$ is the cdf of a normal Gaussian distribution.

The CLT might describes how the distribution of the empirical mean around the theoretic expectation. It might also be surprising since it implies that the mean of discrete random variables could be a continuous random variable!

Exercise 10 : *Find the distribution of the random variable* $\frac{1}{n} \sum_{i=1}^{n} X_i$.

6 Distances and divergences

6.1 Kullback-Leibler divergence

In the following, when introducing information measures, and more specifically, entropy and mutual information, we will need to resort to a measure of distance between two distribution probabilities. There exist many distance measures between probability distributions, but the one which will be of most use to us, will the Kullback Leibler divergence.

Definition 13 (Kullback-Leibler (KL) divergence)

Let P_X and Q_X be two probability distributions defined on a support set $\mathcal{X}$, such that

$$\sum_{x \in \mathcal{X}} P_X(x) = \sum_{x \in \mathcal{X}} P_X(x) = 1.$$

The KL divergence between P_X and Q_X is defined as

$$D_{KL}(P_X || Q_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) \log \left(\frac{P_X(x)}{Q_X(x)} \right). \tag{43}$$

KL-divergence is defined only between laws which share the same support set.

Properties 9 *The KL divergence verifies a certain set of properties :*

1. *Asymmetry :* $D_{KL}(P_X||Q_X) \neq D_{KL}(Q_X||P_X)$.
2. *Null element :* $D_{KL}(P_X||P_X) = 0$.
3. *Positivity :* $D_{KL}(P_X||Q_X) \geq 0$, *for all laws* (P_X, Q_X).

Due to the fact that the KL-divergence is asymmetric, it is denoted as a divergence and not as a distance.

KL-divergence however allows to assess the distance between two distributions, in the sense that it is equal to 0 only when P_X and Q_X are equal, and thus, it allows to test whether two distribution probabilities are close enough to one another.

Exercise 11 *: Prove the properties of the KL divergence. Hint : for the positivity, use the convexity of the log function, or use a Lagrangian to find the minimum of the KL divergence, over all* Q_X *with a* P_X *fixed.*

Conclusions 1 *After this part of the course, you should be able to :*

- *List main discrete and continuous probability distributions*
- *Compute their expectation and variance*
- *Distinguish joint, conditional, and marginal pmfs /pdfs*
- *Enunciate the chain rule, and its applications*
- *List some main results of statistics*
- *Describe the practical implications of these results*