Probability and Statistics

Data Science Engineering

Chapter 1: Probability Spaces

Probability spaces: Random phenomena. Algebra of Events. Probability spaces. Independence. Conditional probability. Bayes Theorem.

As opposite to deterministic experiences, where initial conditions determine the outcome, in random phenomena not one but a collection of outcomes is possible. Randomness is associated to the fact that, in spite of its unpredictability, there is a regularity on the frequency of appearance of these outcomes as the experience is repeated a large number of times. The probability of an event is then reflecting the limiting value of the frequency of its appearance for a large number of repetitions of the same experience. As such, the probability is a measure of chance. The mathematical formalization of this notion from empirical intuition turns out to present some difficulties. Instead, an axiomatic definition is given by keeping in mind that it should reflect the main and basic properties of frequencies of events.

1. Probability Space

The possible outcomes of a random phenomena are identified with subsets of a ground set, usually denoted by Ω and called the *Sample Space*. In this way one can describe complex events in terms of Boolean expressions of simpler events by means of unions, intersections and negation.

A probability is a function on the events (subsets) of the sample space Ω which intends to reflect the chances that a particular event happens as an outcome of the sample space. The measure is normalized to take real values between 0 and 1, and as measure, it should satisfy the additive property for disjoint events. This motivates the following definitions:

Definition 1.1. A sample space associated to an experiment is the set Ω of all possible outcomes. The subsets of Ω are called events.

With the translation of 'events' to 'subsets', every proposition on events can be expressed as a combination of unions, intersections and complements of sets. A review of the basic properties of the Boole algebra of the subsets of a set under these operations is in order. Besides the most obvious properties, like $A \cup \Omega = \Omega$ or $A \cap \emptyset = \emptyset$, one should recall

- Distributivity: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, and
- De Morgan Laws: $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$ and $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$, where here $\overline{X} = \Omega \setminus X$ denotes the complement of a set.

Example 1.2. By throwing a dice, the set of outcomes can be identified by $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event that the result is an even number is $A = \{2, 4, 6\}$ and the event that it is not larger than three is $B = \{1, 2, 3\}$. One can describe the events 'even and smaller than three' as $A \cap B$ and 'even but not smaller than three' by $A \cap \overline{B}$.

There are some subtleties in the definition of general probability spaces which can be avoided in the simpler case of finite probability spaces.

Definition 1.3. A finite probability space is a triple $(\Omega, 2^\Omega, \mathbb{P})$, where Ω is a finite set, 2^Ω is the family of all subsets of Ω and $\mathbb{P} : 2^\Omega \to \mathbb{R}$ is a function taking values on subsets of Ω such that

(1) **non-negativity**: $\mathbb{P}(A) \geq 0$ for each $A \subset \Omega$
(2) **normalization**: $\mathbb{P}(\Omega) = 1$, and
(3) **additivity**: for two disjoint subsets $A, B \subset \Omega$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

The above definition has some simple important implications:

Proposition 1.4. *Let $(\Omega, 2^\Omega, \mathbb{P})$ be a finite probability space.*

(i) *If $A \subset B$ then $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$ and $\mathbb{P}(A) \leq \mathbb{P}(B)$.*
(ii) *For every subset $A \subset \Omega$, $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$.*
(iii) $\mathbb{P}(\emptyset) = 0$.
(iv) $\mathbb{P}(A) \leq 1$, *for $A \subset \Omega$.*
(v) *For every two subsets $A, B \subset \Omega$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.*
(vi) *For pairwise disjoint sets $A_1, \ldots, A_k \subset \Omega$, we have $\mathbb{P}(\cup_{i=1}^k A_i) = \sum_{i=1}^k \mathbb{P}(A_i)$.*

These fairly simple facts allows one to start computing some probabilities in simple situations (coin tossing, dice rolling, etc.)

The probability of the union of several events when they are not pairwise disjoint is less simple to express, but the following is a useful formula:

Proposition 1.5 (Inclusion–Exclusion)**.** *Let $A_1, \ldots, A_k$ be events in a probability space. The probability of the union is*

$$\mathbb{P}\left(\cup_{i=1}^k A_i\right) = \sum_{i=1}^k (-1)^{i+1} \sum_{\substack{S \subset [k] \\ |S| = k}} \mathbb{P}(\cap_{j \in S} A_j).$$

Example 1.6. Choose an integer from 1 to 100. The sample space is $\Omega = [100]$ and all integers have the same probability $1/100$. The probability that it is not divisible by a prime at most 5 can be computed as follows. Let A_i be the event that the chosen integer is not divisible by i. Then we seek for

$$\mathbb{P}(\overline{A_2} \cap \overline{A_3} \cap \overline{A_5}) = \mathbb{P}(\overline{A_2 \cup A_3 \cup A_5}) = 1 - \mathbb{P}(A_2 \cup A_3 \cup A_5),$$

and

$$\begin{aligned}\mathbb{P}(A_2 \cup A_3 \cup A_5) &= \mathbb{P}(A_2) + \mathbb{P}(A_3) + \mathbb{P}(A_5) \\ &\quad - \mathbb{P}(A_2 \cap A_3) - \mathbb{P}(A_2 \cap A_5) - \mathbb{P}(A_3 \cup A_5) \\ &\quad + \mathbb{P}(A_2 \cap A_3 \cap A_5) \\ &= \frac{50}{100} + \frac{33}{100} + \frac{20}{100} - \frac{16}{100} - \frac{10}{100} - \frac{6}{100} + \frac{3}{100} \\ &= \frac{74}{100},\end{aligned}$$

so the probability is $26/100$. $\square$

2. Uniform probability spaces

One simple but common and natural example of finite probability space consists in assigning to each element of Ω the same probability (a uniform probability distribution). If Ω contains n elements, $\Omega = \{\omega_1, \dots, \omega_n\}$, then $\mathbb{P}(\omega_1) = \cdots = \mathbb{P}(\omega_n) = \frac{1}{n}$. In this situation, for every $A \subset \Omega$ we have

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

Therefore, computation of probabilities reduces to counting. Counting is not always a simple task! Here you will have to remember the part of Combinatorics you did in *Lògica i Matemàtica Discreta*. Some important facts that will be useful are:

- **Words**: There are n^k words of length k over an alphabet of n letters. For instance, there are $2^3 = 8$ binary words of length 3:

  $$000, 100, 010, 001, 011, 101, 110, 111.$$

  Example 2.1. Consider a random binary sequence of length k. The sample space can be identified with $\Omega = \{0,1\}^k$ and every string $(x_1, \dots, x_k) \in \{0,1\}^k$ of length k has the same probability $1/2^k$.

  (1) The probability that the string starts with 111 is $1/8$: there are 2^{k-3} strings starting with 111.
  (2) The probability that the string contains exactly three 1's is $\binom{k}{3}/2^k$: there are $\binom{k}{3}$ of them.
  (3) The probability that the sequence starts and finishes with 111 ($k \geq 6$) is $1/64$: there are 2^{k-6} such sequences.

(4) The probability that the sequence does not start with 111 and does not finish with 111 is (if $k \geq 6$) 49/64: if A is the event that the string starts with 111 and B the one that it ends with 111, the probability is

$$\mathbb{P}(\bar{A} \cap \bar{B}) = \mathbb{P}(\overline{A \cup B}) = 1 - \mathbb{P}(A \cup B) = 1 - (\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)) = 1 - (\frac{1}{8} + \frac{1}{8} - \frac{1}{64}).$$

$\square$

- **Permutations**: There are $n!$ permutations of a set with n elements. For instance, one can produce the 6 permutations

$$123, 132, 213, 231, 312, 321$$

of three elements. The function $f(n) = n!$ grows extremely fast (try 10! in a calculator). Its value is well approximated by the Stirling formula $n! \simeq (n/e)^n \sqrt{2\pi n}$.

If we look at 'truncated' permutations, namely, we use only strings of length $k < n$ with pairwise distinct elements, its number is $(n)_k = n(n-1)\cdots(n-k+1)$. For instance, there are 12 permutations of length 2 on 4 symbols:

$$12, 13, 14, 21, 23, 24, 31, 32, 34, 41, 42, 43.$$

Example 2.2. The worst case in the number of steps in a sorting algorithm of n distinct numbers occurs precisely when the numbers appear in decreasing order. If all orderings are equally possible, the sample space Ω consists of all permutations of the numbers and each permutation has the same probability. Thus, the probability of having a worst case instance is $1/n!$. $\square$

- **Subsets**: The number of subsets of cardinality k from a set of cardinality n is the *binomial coefficient*

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

This is one of the most important family of numbers and one should understand well its meaning and applications.

Some important properties are:

(i) $\binom{n}{0} = \binom{n}{n} = 1$.

(ii) $\binom{n}{k} = \binom{n}{n-k}$.

(iii) $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}, n > k \geq 1$.

(iv) $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

Example 2.3. In a set of n people a commission of k representatives is chosen. If all selections are equally liking, the sample space is the set of all k–subsets of $[n]$ and each one has the same probability $1/\binom{n}{k}$. The number of k–subsets containing one particular person is $\binom{n-1}{k-1}$, so that the probability of the event $A = \{\text{Anna is chosen}\}$ is

$$\mathbb{P}(A) = \frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}.$$

$\square$

- **Sampling**: Selection of a sample of k individuals from a population of n individuals. The number of samples depends on the sampling procedure and the criteria to distinguish different samples. The sampling can be done with or without replacement. Moreover, the ordering in which individuals are selected can be taken into account or not. According to the different types of sampling, the number of samples is

number of samples	Ordered	Unordered
replacement	n^k	$\binom{n+k-1}{k}$
no replacement	$(n)_k$	$\binom{n}{k}$

Example 2.4. We pick five cards in a french deck of 52 cards. The probability of getting a trio is

$$\frac{\binom{13}{1}\binom{4}{3}\binom{48}{2}}{\binom{52}{5}} \approx 0.023$$

where the denominator is the total number of possible hands of 5 cards, and the numerator counts the number of them containing a trio: choose one number, three cards of this number and the two remaining ones are free but distinct from the ones in the trio (a full is possible). □

3. Independence

Stochastic independence is a crucial notion in probability theory. Given events $A, B \subset \Omega$, the fact that one occurs can affect the probability that the other one does. From the mathematical perspective, this is measured by comparing $\mathbb{P}(A \cap B)$ and $\mathbb{P}(A)\mathbb{P}(B)$. For independent events, the seemingly natural equality $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ holds. This is in fact the mathematical definition of independence.

Definition 3.1. Two events $A, B \subset \Omega$ are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

A family $A_1, \dots, A_k$ of events is *mutually independent* if for every subset $S \subset [k]$,

$$\mathbb{P}(\cap_{i\in S} A_i) = \prod_{i\in S} \mathbb{P}(A_i).$$

Independence is not to be confused with incompatibility. We say that two events $A, B \subset \Omega$ are incompatible if $\mathbb{P}(A \cap B) = 0$.

Example 3.2. We toss a fair coin k times and consider the sequence of outcomes. We denote by A_i the event in the i–th toss is a Head. The sample space can be identified with $\Omega = \{0, 1\}^k$ (say 0 is Heads and 1 is Tails) and it has $|\Omega| = 2^k$ elements. The underlying assumption is that all elements in the sample space have the same probability. Then the probability that $A_1 = 0$ is

$$\mathbb{P}(A_1) = \frac{1}{2},$$

(half of the sequences start with 0), the probability that $A_2 = 0$ is also

$$\mathbb{P}(A_2) = \frac{1}{2},$$

and the probability that the first two outcomes are zero is

$$\mathbb{P}(A_1 \cap A_2) = \frac{1}{4},$$

(there are 2^{k-2} sequences starting with 00.)

Therefore, the assumption that all outcomes are equally likely leads to the outcomes in the first and in the second toss being independent. More generally, if we select two disjoint sets of subscripts, that is $I, J \subset [k]$ with $I \cap J = \emptyset$ (for instance, $I = \{1, 3, 4\}$ and $J = \{2, 5, 6\}$) the events $A_I = \cap_{i \in I} A_i$ and $A_J = \cap_{j \in J} A_j$ are also independent. In our example, $\mathbb{P}(A_I) = \mathbb{P}(A_J) = 1/8$ and $\mathbb{P}(A_I \cap A_J) = 1/64 = 1/8\dot{1}/8 = \mathbb{P}(A_I)\mathbb{P}(A_J)$. In particular, the family $A_1, \ldots, A_k$ is mutually independent.

Reciprocally, the assumption that the outcomes in different tosses are mutually independent (an intuitively reasonable assumption) leads to the probability of each sequence being $1/2^k$, so that they are all equally likely.

One can for instance deduce that the probability of finding the sequence 0010 in four consecutive positions is not smaller than $1 - (1 - 1/16)^{\lceil k/4 \rceil}$: $(1 - 1/16)$ is the probability of not finding that particular string in the sequence starting at some position. The events of finding that string starting in different positions are independent if the positions are at least four units apart, and we can find $\lceil k/4 \rceil$ such positions. We observe that this probability tends to 0 when k goes to infinity, so for large k the string is very likely to appear in the sequence; this is a particular instance of the so-called *Infinite monkey theorem*.

In the above problem, computing exactly the probability of not finding 0010 is possible, but it requires considerably more work.

One example of non mutually independent events which are however pairwise independent is as follows. Let $A_{ij} = \{x \in \Omega : x_i = x_j\}$. If $i \neq j$, $\mathbb{P}(A_{ij}) = 1/2$. Moreover, $\mathbb{P}(A_{12} \cap A_{13}) = \mathbb{P}(A_{12} \cap A_{23}) = \mathbb{P}(A_{13} \cap A_{23}) = 1/4$, so the three events are pairwise independent, but $\mathbb{P}(A_{12} \cap A_{13} \cap A_{23}) = 1/4 \neq \mathbb{P}(A_{12})\mathbb{P}(A_{13})\mathbb{P}(A_{23})$, so the there events are not mutually independent.

Another example of non joint independence is the following one: take $k = 12$ and choose one position x at random. Let $C_1 = \{x \leq 6\}, C_2 = \{6 \leq x \leq 9\}$ and $C_3 = \{x \in \{4, 5, 6, 10, 11, 12\}\}$. We have $\mathbb{P}(C_1) = \mathbb{P}(C_3) = 1/2$ and $\mathbb{P}(C_2) = 1/3$. The events are not independent since $\mathbb{P}(C_1 \cap C_2) = 1/12 \neq \mathbb{P}(C_1)\mathbb{P}(C_2)$ even if $\mathbb{P}(C_1 \cap C_2 \cap C_3) = 1/12 = \mathbb{P}(C_1)\mathbb{P}(C_2\mathbb{P}(C_3)$. $\square$

Example 3.3 (Checking polynomial identities)**.** We have two polynomials $P_1(x), P_2(x)$ of degree d and wish to know if $P_1(x) = P_2(x)$ (say $P_1(x) = (x-3)(x^2+5x-1)(x^4+2)$ and $P_2(x) = x^7 + (x^6-1)(x^2+2x-3)$).

A simple algorithm is to select at random a number r in $[100d]$ and evaluate $P_1(r)$ and $P_2(r)$. If the evaluation gives a different answer, then certainly $P_1 \neq P_2$. The evaluation may give by chance the same answer and still the polynomials be distinct. In that case, r is a root of $P_1(x) - P_2(x)$, so the probability of this happening is $1/100$ (there are only d roots). The probability is quite small, so we may output that the polynomials are the same if $P_1(r) = P_2(r)$.

If we repeat the algorithm k times independently, then the probability that we make an error (namely, the values coincide the k times but $P_1 \neq P_2$) is $(1/100)^k$, perhaps smaller than the probability of an error in computing the standard form of the two polynomials, and certainly can be made arbitarily small, with a very simple algorithm. One could certainly evaluate the polynomials in $d+1$ points to have a deterministic algorithm. But if d is large, the probabilistic version gives a reasonable probability of error with smaller number of evaluations. This is a clever use of probability. $\square$

4. Conditional Probability and Bayes theorem

When two events A, B are not independent the appearance of A may affect the probability of the appearance of B by increasing or decreasing its probability. The natural way to express the degree and nature of dependence is through the *conditional* probability.

Definition 4.1. Let B be an event with $\mathbb{P}(B) > 0$. The conditional probability of an event A given B is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

The function $\mathbb{P}(\cdot|B) : 2^{\Omega} \to [0,1]$ defined as in (1) is again a probability in the same sample space B.

Equality (1) can also be read off as

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B), \tag{2}$$

giving an expression for the probability of intersection which, of course, depends on the nature of the dependence of A with B. In particular, the two events A, B are independent if and only if

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

which links the notion of independence to its intuitive meaning: knowing that B holds does not affect the probability A holds.

Example 4.2. An urn contains three white balls and two black ones. We sample two balls without replacement. The probability that they are both white is

$$\mathbb{P}(\circ\circ) = \mathbb{P}(* \circ \mid \circ *)\mathbb{P}(\circ *) = (1/2) \cdot (3/5) = 3/10,$$

where here the event A that the first ball is white is represented as '$\circ *$' and the event B that the second ball is white is '$* \circ$', so that $A \cap B$ is written '$\circ\circ$'. $\square$

Equation (2) can be extended to longer intersections:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B|A)\mathbb{P}(C|A \cap B). \tag{3}$$

Example 4.3. An urn contains three white balls and two black ones. We sample three balls without replacement. The probability that the three are white is

$$\mathbb{P}(\circ \circ \circ) = \mathbb{P}(\circ * *)\mathbb{P}(\circ \circ *| \circ **)\mathbb{P}(\circ \circ \circ| \circ \circ *) = (3/5) \cdot (2/4) \cdot (1/3) = 1/10.$$

$\square$

It is often the case that conditional probabilities are easier to obtain than unconditional ones. The so–called *Total Probability Theorem (TPT)* is frequently used. A *partition* of the sample space Ω is a sequence of sets $A_1, \dots, A_k$ such that $\cup_{i=1}^k A_i = \Omega$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Theorem 4.4 (TPT)**.** *Let A, B be events,*

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\overline{A})\mathbb{P}(\overline{A}).$$

More generally, if $A_1, \dots, A_k$ is a partition of Ω then

$$\mathbb{P}(B) = \mathbb{P}(B|A_1)\mathbb{P}(A_1) + \cdots + \mathbb{P}(B|A_k)\mathbb{P}(A_k).$$

Example 4.5. An urn contains three white balls and two black ones. We sample two balls without replacement. The probability that the second ball is white is

$$\begin{aligned}\mathbb{P}(*\circ) &= \mathbb{P}(* \circ \mid \circ *)\mathbb{P}(\circ *) + \mathbb{P}(* \circ \mid \bullet *)\mathbb{P}(\bullet *)\\ &= (1/2) \cdot (3/5) + (3/4) \cdot (2/5) = 3/5.\end{aligned}$$

the same as the first ball is white. $\square$

The formula for the conditional probability is kind of symmetric in A and B in the sense that, if $\mathbb{P}(A) \neq 0$ and $\mathbb{P}(B) \neq 0$ then

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

This observation is usually applied in combination with the formula of total probability to relate $\mathbb{P}(A|B)$ with $\mathbb{P}(B|A)$ in the so–called *Bayes Theorem.*

Theorem 4.6. *Let $A, B \subset \Omega$ with $\mathbb{P}(A) \neq 0$ and $\mathbb{P}(B) \neq 0$, then*

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

Moreover, if $A_1, \ldots, A_k$ is a partition of Ω, then for all $i \in [k]$

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B|A_1)\mathbb{P}(A_1) + \cdots + \mathbb{P}(B|A_k)\mathbb{P}(A_k)}.$$

Example 4.7. An urn contains three white balls and two black ones. We sample two balls without replacement. The probability that the fist ball is white if the second is white is

$$\begin{aligned}\mathbb{P}(\circ * \mid * \circ) &= \frac{\mathbb{P}(* \circ \mid \circ *)\mathbb{P}(\circ *)}{\mathbb{P}(* \circ \mid \circ *)\mathbb{P}(\circ *) + \mathbb{P}(* \circ \mid \bullet *)\mathbb{P}(\bullet *)} \\ &= \frac{(1/2) \cdot (3/5)}{3/5} = 1/2.\end{aligned}$$

□

5. Infinite sample spaces

So far we have only considered finite sample spaces. Infinite sample spaces occur naturally, countable or uncountable ones. Simple examples include

- the first time a head shows up in tossing a coin;
- the analysis of infinite binary sequences, which are in correspondence with the real interval $[0, 1]$ (via their binary expression);
- the time when a random event occurs (first arrival in a queue).

It turns out that our axiomatic definition of probability spaces can be adapted in a simple way to the infinite setting, but some language must be introduced.

Definition 5.1 (σ–algebra of events)**.** Let Ω be a set. A family $\mathcal{A} \subset 2^{\Omega}$ of subsets of Ω is a *σ–algebra* on Ω if

(1) $\Omega \in \mathcal{A}$,
(2) if $A \in \mathcal{A}$, then $\overline{A} \in \mathcal{A}$, and
(3) $A_1, A_2, \ldots$ a sequence of events each one in $\mathcal{A}$, then $\cup_{i \geq 1} A_i \in \mathcal{A}$.

A set $A \in \mathcal{A}$ is called an event, and we say that it is *measurable* with respect to $\mathcal{A}$.

So a σ–algebra is a collection of subsets closed under complementation and countable unions. By the De Morgan law, the σ-àlgebra is also closed under countable intersections.

Example 5.2. Simple examples of σ–algebras are

(1) $\mathcal{A} = 2^{\Omega}$ (complete σ-algebra, also called power-set if $\mathcal{A}$ is finite).

(2) $\mathcal{A} = \{\emptyset, \Omega\}$ (trivial σ-algebra).
(3) $\mathcal{A} = \{\emptyset, \mathcal{A}, \overline{A}, \Omega\}$, where $A \subset \Omega$ (the Bernouilli σ–algebra).

The reason to introduce this notion is twofold. On one hand, one may not be interested in all possible events, but only a subfamily of them. From the other one, there is no way to define a probability for all subsets of the real interval $[0, 1]$. This is a deep result in Measure Theory that we will not address. However, a probability can be defined for a reasonably wide class of subsets forming a σ–algebra.

Definition 5.3. Let $\mathcal{A}$ be a family of subsets of a set Ω. The σ–algebra $\sigma(\mathcal{A})$ generated by $\mathcal{A}$ is the smallest σ–algebra that contains $\mathcal{A}$.

For example, if A is a single proper subset of Ω then $\sigma(\{A\}) = \{\emptyset, A, \bar{A}, \Omega\}$.

Example 5.4. Let $\Omega = 1, 2, 3, 4, 5$ and $\mathcal{A} = \{\{1, 2\}, \{3\}\}$, then

$$\sigma(\mathcal{A}) = \{\emptyset, \{1, 2\}, \{3\}, \{1, 2, 3\}, \{3, 4, 5\}, \{1, 2, 4, 5\}, \{4, 5\}, \Omega\}$$

For instance, $\{1, 3, 4\} \notin \sigma(\mathcal{A})$.

Example 5.5. We toss a coin three times. Our sample space is $\{0, 1\}^3$. Then, we can differentiate 4 σ-algebras that appear sequentially in the experiment:

- $\mathcal{A}_0 = \{\emptyset, \Omega\}$, the trivial σ-algebra before starting the experiment.
- $\mathcal{A}_1 = \{\emptyset, \{000, 001, 010, 011\}, \{100, 101, 110, 111\}, \Omega\}$, the σ-algebra obtained after tossing the first coin. That is, after tossing the first coin we cannot distinguish between experiments that have the same initial outcome.
- $\mathcal{A}_2 = \sigma(\{000, 001\}, \{010, 011\}, \{100, 101\}, \{110, 111\}\})$, the σ-algebra obtained after the second toss is generated by the sets of elements that coincide in the first two positions.
- $\mathcal{A}_3 = 2^\Omega$, after the third toss, all the information is revealed and we obtain the power-set.

Definition 5.6. The Borel σ–algebra $\mathcal{B}(\mathbb{R})$ of $\mathbb{R}$ is the σ–algebra generated by all the open intervals $(a, b) \subset \mathbb{R}$.

It can be shown that $\mathcal{B}(\mathbb{R})$ does not contain all subsets of $\mathbb{R}$, but it contains all reasonable sets from the point of view of probability.

We can extend the notion of probability spaces to include the infinite ones.

Definition 5.7 (Probability space)**.** A probability space is a triple $(\Omega, \mathcal{A}, \mathbb{P})$ where Ω is a set, $\mathcal{A}$ is a σ–algebra of subsets of Ω and $\mathbb{P} : \mathcal{A} \to \mathbb{R}$ is a function satisfying

(1) **non-negativity**: $\mathbb{P}(A) \geq 0$ for each $A \subset \Omega$
(2) **normalization**: $\mathbb{P}(\Omega) = 1$, and

(3) **additivity**: if $A_1, A_2, \ldots$ is a sequence of pairwise disjoint subsets in $\mathcal{A}$, then

$$\mathbb{P}(\cup_{i\geq 1} A_i) = \sum_{i\geq 1} \mathbb{P}(A_i).$$

This notion includes our previous one for finite sample spaces.