1
Probability

> The Probability of an Event is greater, or less, according to the number of Chances by which it may Happen, camper'd with the number of all the Chances, by which it may either Happen or Fail.
>
> The Doctrine of Chance,
> Abraham de Moivre

1.1 Randomness and Probability

What are the odds that the the final match of the World Cup in 2022 is played by Spain and Germany? How likely is it for the Euro/USD exchange rate to fall by 20 percent by the end of 2022? Compare these questions with the following questions: How long does it take for the light to arrive from Alpha Centauri to earth? What will be the velocity of a Euro coin after one minute when dropped in vacuum at a specific point on earth? We tend to believe that, just as many events in the physical world are governed by deterministic natural laws—think of the Newton's laws of motion governing the motion of celestial objects— and hence, can be predicted with a good degree of accuracy, for many other events, such laws are not available or, at least, not easy to formulate. When a coin is tossed, in the absence of more information about how it is tossed, we can only talk about the set of different outcomes. In such cases, we also tend to ascribe odds, likelihood, or a probability to each possible outcome to quantify our degree of belief in the occurrence of the each outcomes.

The roots of the probability theory at least go back to the seventeenth century, when it was used as a tool for solving a variety of practical problem. Analyzing games of chance, making aleatory contracts, and many other practical problems naturally lead to the questions that can only be answered by a theory that quantifies probability, risk, and related concepts.

The modern probability theory is the mathematical framework for systematically quantifying and studying uncertainty that was established by by Andrey Kolmogorov in his groundbreaking work *Grundbegriffe der Wahrscheinlichkeitsrechnung* published in 1933. Kolmogorov laid down the

foundations of probability theory on the measure theory that has been developed in 19th and 20th century by mathematicians such as Camille Jordan, Émile Borel and Henry Lebesgue.

1.2 *Probability as Frequency*

One of the ubiquitous objects in probability theory is the so-called "fair coin". A fair coin or an ideal coin is, by definition, a coin with the property that lands heads or tails with equal probability $1/2$. Such a coin can be used as a tie-breaking tool in Assuming that these are the only possible outcomes (for instance, the coin does not land on the side, or disappear in midair), each of the two events must have probability $1/2$. The definition looks deceptively simple and straightforward. But once we try to understand it better, we will start to observe its shortcomings. To start the discussion, let us see how one can determine if a given coin is fair or not. Can one design a process that determines in finite time whether a given physical coin is fair or not? Obvious attempts at coming up with a criterion quickly fail. For instance, let us define a coin to be fair when it passes the following test: when the coin is dropped $2n$ times (for a large value of n) it lands n times head and n times tail. Such a definition is bad, because even for an idea fair coin, the probability of getting n heads and n tails in $2n$ throws is very small (see Remark **??**). A better attempt is to define a coin fair if

$$\lim_{n\to\infty} \frac{h_n}{n} = 1/2,$$

where h_n is the number of heads in the first n throws. As reasonable as it may sound, this is not a useful definition. To see this, note that the question of whether a sequence a_n converges to a limit a or not cannot be answered by examining a finite number of the terms of the sequence. In fact, for a fair coin, there is a positive, even though tiny, probability that the first 1000 throws result in 1000 heads. So, obtaining 1000 straight heads in 1000 throws does not automatically rule out the possibility that the coin is fair.

The above discussion reveals some of the difficulties that one faces in developing a theory of probability that is based on an objective notion such a frequency. In what follows, we will see that one can do away with an objective theory by interpreting the probabilities as the degree of belief in a possible outcome. Such a theory has the added benefit that it also provides justification in using Kolmogorov's formalism of probability theory in dealing with the everyday problems.

1.3 *Subjective Probability of Ramsey-De Finetti*

In developing a theory of probability, an alternative approach to is to view probability as the degree of certainty or belief in the occurrence of a possible outcome. This degree of belief is quantified by a number between 0 and 1. Here, 1 will correspond to a certain outcome, while 0 corresponds to an outcome that can be ruled out. Any number in between represents the degree of likelihood in between.

Let us make the discussion more concrete by imagining a company that announces (explicitly or implicitly) a probability for possible outcomes of a game. The word "game" must be understood in an abstract sense: A company that sells insurance contracts or an online gambling

website, where the participants can bet on the possible outcomes of various athletic or political events can be viewed as examples.

To facilitate the discussion, we will assume that the set of all possible outcomes of an event is a finite set given by

$$O = \{o_1, \ldots, o_n\}.$$

The company also allows placing bets for or against any set of possible outcomes. More precisely, for any subset $A \subseteq O$ of the outcomes, the company announces a probability $\mathbb{P}[A]$ which stands for company's estimate of the odds that the outcome of the event is in A. The company is then ready to accepts bets that are consistent with the announced odds.

For instance assume that the $O = \{o_1, o_2, \ldots, o_{32}\}$ are the set of possible winners of the FIFA World Cup 2014, where each $o_1, \ldots, o_{32}$ denote the thirty two finalist teams, enumerated in a specific (but, arbitrary) way. Assume that

$$A = \{o_1, o_7, o_{12}\},$$

and the company assigns the odds $p_A = 2/17$ to A. This means, that you can bet on the winner being in A in which case, you can pay 1 euro upfront, and if the winner of the World Cup turns out to be $o_1, 0_7$, or o_{12}, you will win $1/p_A = 17/2$ euros. Of course, if you believe that the company overestimates the probability, you can place a bet against the company by agreeing to pay the company $1/p = 17/2$ euros in exchange for receiving a sure amount of 1 euro now. Thus, the company has assigned a number p_A to every subset $A \subseteq O$, standing for the probability that the outcome of the game is in A. How arbitrary can these numbers be?

Let us see how an assignment of probabilities can be incoherent. Suppose, in the above example, the company assigns a higher probability $p_B = 1/8$ to a proper subset

$$B = \{o_1, o_7\} \subset A = \{o_1, o_7, o_{12}\}.$$

Regardless of what how ones thinks about the chances of the teams o_1, o_7, o_{12} to win the championship, it is possible to place some bets in such a way that regardless of the outcome of the game, the company will be beaten.

In order to make the discussion more precise, let us say that an assignment of probabilities $A \mapsto \mathbb{P}[A]$ is incoherent if it is possible to place a number of bets in such a way that no matter how the outcome of the event turns out to be, the total sum of money that the company pays is non-negative, there exists an outcome for which the company payoff is positive.

Proposition 1.3.1. *For any coherent assignment of probabilities, we have* $\mathbb{P}[O] = 1$.

Proof. If the company assigned $\mathbb{P}[O] = 1 + a > 1$, the obviously one buys a ticket at the cost of 1 Euro which pays $1 + a$ if the outcome lands in O. But this is the set of all possible outcome. So the company loses $a > 0$ Euros regardless of the outcome of the game. If $\mathbb{P}[O] = 1 - a < 1$, one can make the opposite bet. □

Proposition 1.3.2. *In any coherent assignment of probabilities, we have* $\mathbb{P}[A] + \mathbb{P}[O \setminus A] = 1$.

Proof. Let us assume that $\mathbb{P}[A] = a$, $\mathbb{P}[O \setminus A] = b$ and $a + b < 1$. Let us make a bet on both A and $O \setminus A$. The cost of doing so is $a + b$. On the other hand, regardless of the outcome of the game we will be paid the constant sum of 1 Euros, hence we can guarantee the sum of $1 - a - b$. □

Similarly, we can show the following property:

Proposition 1.3.3. *If A and B are disjoint subsets of O, then*

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$$

Proof. Let us assume that $\mathbb{P}[A \cup B] > \mathbb{P}[A] + \mathbb{P}[B]$. We will make three bets: one against $A \cup B$, one for A and one for B. Note that our balance after making these bets is $\mathbb{P}[A \cup B] - \mathbb{P}[A] - \mathbb{P}[B] > 0$. Once again, regardless of what happens, these transactions pay for themselves after the end of the game. If the outcome is in A, then we win 1 Euros and lose 1 (because it is also in $A \cup B$). By the same token, if the outcome is in B, we lose 1, but that will be paid by the 1 Euro we win from the bet on $A \cup B$.

□

1.4 Axiomatization of Finite Probability Spaces

In mathematical probability we always start with a finite set that is usually denoted by Ω and is called the *sample space*.[1] A subset of Ω is called an *event*. Simply put, the idea is that the each element of the sample space stands for a possible outcome of an experiment where we cannot predict the result in advance (e.g. flipping a coin or throwing a die), and an event is a set of possible outcomes. Of course, the outcomes may not be equally likely. The likelihood is measured by what we call the probability function assigned to the events. This is formulated in the following definition.

[1] Ω (read omega) is Greek alphabet.

Definition 1.4.1. Let Ω be a non-empty set. A function $\mathbb{P}$ that assigns to every subset $A \subseteq \Omega$ a number $\mathbb{P}(A)$ is called a *probability map* if

(i) $\mathbb{P}[A] \geq 0$, for all $A \subseteq \Omega$;

(ii) $\mathbb{P}[\Omega] = 1$;

(iii) $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$, given that $A \cap B = \varnothing$.

Remark 1.4.2. The axioms are chosen to reflect what we expect from the probability to measure. The empty set representing the event that none of the possible outcomes occur has zero probability. We also need the third axiom to ensure that if two events are disjoint, then the probability

that at least one of them occurs is the sum of the corresponding probabilities. The second axiom can be viewed as a normalizing axiom.

Here are some of the implications of these axioms.

Theorem 1.4.3. *The probability map has the following properties:*

1. $\mathbb{P}[\varnothing] = 0$.
2. $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$
3. *If* $A \subseteq B$ *then* $\mathbb{P}[A] \leq \mathbb{P}[B]$.
4. *For any* $A,\ B \subseteq \Omega$ *we have*

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$$

Proof. Since $\varnothing \cap \varnothing = \varnothing$ we have $\mathbb{P}(\varnothing \cup \varnothing) = \mathbb{P}(\varnothing) + \mathbb{P}(\varnothing)$ which implies $\mathbb{P}(\varnothing) = 0$. For (b), note that $A \cap A^c = \varnothing$ and $A \cup A^c = \Omega$. So $1 = \mathbb{P}[A] + \mathbb{P}[A^c]$. For (c), write

$$B = A \cup (B - A), \quad A \cap (B - A) = \varnothing.$$

So $\mathbb{P}[B] = \mathbb{P}[A] + \mathbb{P}[B \setminus A] \geq \mathbb{P}[A]$. For (d), Since

$$\mathbb{P}[B] = \mathbb{P}[B \cap A] + \mathbb{P}[B \setminus A],$$

we have

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B - A] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B].$$

□

There is a useful generalization of this proposition to the union of more than two sets that will be discussed later.

1.5 Identifying the Sample Space

In any random experiment, the sample space consists of the set of all possible outcomes. The sample space, per se, is not sufficient for a probabilistic analysis. One also needs to have an assignment of probabilities to the subsets of the sample space that is consistent with the axioms of probability. In the next examples, we will identify the sample space for a few typical random experiments.

Example 1.5.1. A pair of coins have been tossed. If we denote the sides of the coins by head and tail, the set of all possible outcomes can be identified with

$$\Omega = \{HH, HT, TH, TT\}.$$

Example 1.5.2. A die is rolled. The sample space in this case equals

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Example 1.5.3. The price of a stock is registered at times t_1, t_2, t_3. Assuming that the price is a positive real number, the sample space is the set

$$\Omega = \{(x_1, x_2, x_3) : x_1, x_2, x_3 > 0\}.$$

If, instead, we register the price of the stock as a function $p(t)$ of time t over the period $[0, T]$, then the sample space consists of all functions $p : [0, T] \to (0, \infty)$. One can, of course, impose more restrictive conditions on $p(t)$. For instance, it is natural to assume that $p(t)$ is a continuous function of t. This will reduce the sample space to the set of all *continuous* functions $p : [0, T] \to (0, \infty)$.

1.6 The Equiprobable Outcomes– Probability and Counting

After identifying the sample space, we turn to the question of assigning probabilities. We will start with the case of finite probability spaces. Let us denote the sample space by $\Omega = \{\omega_1, \ldots, \omega_n\}$, where each ω_j is one of the outcomes. Let us set $p_j = \mathbb{P}\left[\{\omega_j\}\right]$. For each set $A \subseteq \Omega$, the additivity axiom implies that

$$\mathbb{P}[A] = \sum_{\omega_j \in A} p_j.$$

It is also easy to see that if $p_j \geq 0$ and $\sum_{j=1}^{n} p_j = 1$, then the map defined above is a probability map.

An important case is when all the elements of Ω are equiprobable. This implies that for each $\omega \in \Omega$, we must have

$$\mathbb{P}[\{\omega\}] = \frac{1}{|\Omega|}.$$

Now, using axiom (iii) we can see that if $A = \{\omega_1, \ldots, \omega_k\}$, then we have:

$$\mathbb{P}[A] = \mathbb{P}[\{\omega_1\}] + \cdots + \mathbb{P}[\{\omega_k\}] = \frac{k}{n} = \frac{|A|}{|\Omega|}.$$

Definition 1.6.1 (Uniform Probability)**.** Let Ω be a finite set. The uniform probability on Ω is defined by

$$\mathbb{P}[A] = \frac{|A|}{|\Omega|},$$

for every subset $A \subseteq \Omega$.

The discussion above shows that the computing probabilities when the outcomes are equiprobable is in essence an enumeration problem. More precisely, in order to compute the probability of an event A, it suffices (but not necessary; cf. Example **??**) to compute $|A|$ and $|\Omega|$, that is, solve two *counting* problems. Counting a set defined by a set of constraints could be quite involved. We will confine ourselves to some of the most basic techniques.

Example 1.6.1. A 3-digit number is chosen randomly. What is the probability that the sum of the digits of the number is even?

In this example we can describe the sample space as

$$\Omega = \{100, \dots, 999\}.$$

Therefore $|\Omega| = 999 + 1 - 100 = 900$. This can also be obtained using the multiplication principle:

$$|\Omega| = 9 \times 10 \times 10.$$

Now it remains to compute $|A|$. It is clear that the left-most digit can be chosen in 9 and the middle one in 10 ways, respectively. Once these two are chosen, there will be 5 possibilities for the right-most digit: If the sum of the two already chosen numbers is even, the last digit must be also even and if the sum of the chosen numbers is odd, the last digit must be also odd. So, $|A| = 9 \times 10 \times 5$, and

$$\mathbb{P}[A] = \frac{|A|}{|\Omega|} = \frac{9 \times 10 \times 5}{9 \times 10 \times 10} = \frac{1}{2}.$$

Tip. Choosing the approach in counting is essential. For example, try to redo the counting Example **??**, but start counting from the right-most digit. You will see that you will need to take cases and counting will be less straightforward. So, if one method does not work, try other methods.

Example 1.6.2. Consider a sequence of independent tosses of a fair coin. The possible outcomes could be head or tail. Let us denote the possible outcome of the i-th trial by X_i. Here the sample space can be described as

$$\Omega = \{(X_1, \dots, X_n) : \ X_i \in \{H, T\}\}$$

Here H and T stand for head and tail, respectively.

Theorem 1.6.2 (Counting without Order)**.** *If A is a set with n elements then the number of subsets with r elements is*

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

Proof. Using Counting Principle 1

$$\binom{n}{r} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!}\frac{(n-r)!}{(n-r)!} = \frac{n!}{r!(n-r)!}$$

□

Example 1.6.3. A committee of 5 persons is going to be formed from 6 men and 9 women. If all of the possible committees have he same , what is the probability that the committee consists of 3 men and 2 women?

Using the counting principles 1,2 we have:

$$\mathbb{P}[A] = \frac{\binom{6}{3}\binom{9}{2}}{\binom{15}{5}}$$

Example 1.6.4. A closet contains m pairs of shoes. n shoes are chosen at random from the closet. Compute the probability that there will be exactly k complete pair, where $0 \leq k \leq m$.

First note that the probability will be zero if $n > m + k$, as in this case there exists at least $k + 1$ complete pairs. Let us assume that $n \leq m + k$. The number of the ways to choose n shows from m pairs of shoes is given by $\binom{2m}{n}$. Note that we are not taking the order in which the shoes are selected into consideration. The number of desirable selections—those in which there are exactly k complete pairs can be calculated as follows. First we choose k of the pairs that are supposed to be complete. This can be done in $\binom{m}{k}$ ways. This provides $2k$ out of the total of n shoes. Next, out of the remaining $m - k$ pairs, we need to choose $n - 2k$ shoes in such a way that no two shoes belong to the same pair. To this end, we first choose $n - 2k$ of the set of pairs, and then from each pair we choose exactly one shoe. The number of the ways in which this can be done is $\binom{m-k}{n-2k}2^{n-2k}$. Hence, the probability of ending up with exactly k complete pairs is given by

$$\mathbb{P}[A] = \frac{\binom{m}{k}\binom{m-k}{n-2k}2^{n-2k}}{\binom{2m}{n}}.$$

Note that when $n > m + k$, we have, by definition, $\binom{m-k}{n-2k} = 0$, hence the special case singled out at the beginning will be also covered by the general formula.

Example 1.6.5. A group of $2n$ girls and $2n$ boys is randomly divided into two groups of size $2n$. Find the probability p_n that each group consists of n boys and n girls, and shows that as $n \to \infty$, we have

$$p_n \sim \sqrt{\frac{2}{\pi n}}.$$

First notice that any partition corresponds to a pair (A, B) of subsets of the set of all $4n$ people into disjoint subsets A and B of size $2n$. As A determines B, we have $|\Omega| = \binom{4n}{2n}$. Now, let E denote the set of all such divisions in which A consists of n boys and n girls. By the same reasoning, we have

$$|E| = \binom{2n}{n}\binom{2n}{n} = \frac{(2n)!^2}{n!^4}.$$

From here, we have

$$\mathbb{P}[E] = \frac{|E|}{|\Omega|} = \frac{(2n)!^4}{n!^4(4n)!}.$$

Using Stirling's Formula, one can easily find the following asymptotic formula for $\mathbb{P}[E]$:

$$\mathbb{P}[E] \sim \frac{(4\pi n)^2(\frac{2n}{e})^{8n}}{(2\pi n)^2(\frac{n}{e})^{4n}\sqrt{8\pi n}(\frac{4n}{e})^{4n}} = \sqrt{\frac{2}{\pi n}}.$$

Example 1.6.6 (Tipping an election)**.** In a presidential election, with two candidates, what is the probability that your vote tips the election, if the number of voters in $2n+1$? (We assumed that the number is odd to rule out the possibility of a tie).

A vote will tip the election, if the remaining $2n$ votes are split equally between the candidates. For simplicity, let us assume that each one of the votes is cast randomly and with probability $1/2$ for one of the candidates. Then, the probability of this event is

$$\mathbb{P}[A] = \binom{2n}{n}\frac{1}{4^n} \approx \frac{1}{\sqrt{\pi n}},$$

where the approximation is done using Stirling's formula. For instance, if the population of the voters is 1000 then the probability is approximately 0.018.

Example 1.6.7 (*The Birthday Problem)*. 25 students are at a party. What is the probability that two of them are born on the same day of the year?

This situation can be modeled in the following way. To each one of the students associate an integer i from the set $\{1,2,\ldots,25\}$. Associate to each day of the year a number from the set $\{1,2,\ldots,365\}$. Here we are ignoring the leap years. Now we can describe the sample space as

$$\Omega = \{(x_1,x_2,\ldots,x_{25}) : x_i \in \mathbb{Z}, 1 \leq x_i \leq 365\}.$$

Here x_i denotes the day in the year in which students with number i is born. So $|\Omega| = 365^{25}$. We can describe the event of interest as

$$A = \{(x_1,x_2,\ldots,x_{25})|\ x_i = x_j \text{ for some } i \neq j\}.$$

Here is the main trick in this problem. Instead of counting A directly (which is not that easy) we will count the complement event A^c. You will see that it is indeed easier. Note that

$$A^c = \{(x_1,x_2,\ldots,x_{25})|\ x_i \neq x_j \text{ for each } i \neq j\}.$$

So using MP we have $|A^c| = 365\cdot(365-1)\cdots(365-24)$ and

$$\mathbb{P}[A^c] = \frac{365\cdot(365-1)\cdots(365-24)}{365^{25}} = (1-\frac{1}{365})(1-\frac{2}{365})\cdots(1-\frac{24}{365}) \approx 0.431.$$

Therefore $\mathbb{P}[A] \approx 0.57 > 1/2$ which may come as a surprise, given that the number of students is much less than the number of the days in a year.

Here is a variation of the birthday problem in a more general setting:

Example 1.6.8. Let A be a set with m elements and B a set with n elements and F denote the set of all functions from A to B.

1. Show that $|F| = n^m$.
2. A function $f \in F$ is randomly drawn. Let $p(m,n)$ be the probability of the event that f is one-to-one, i.e. $f(x) = f(y)$ for $x,y \in A$ implies that $x = y$. Show that

$$p(m,n) = \begin{cases} \prod_{j=1}^{m-1}\left(1-\frac{j}{n}\right) & \text{if } m \leq n \\ 0 & \text{otherwise} \end{cases}$$

3. Show that for any real number $0 \leq x \leq 1$, we have $1 - x \leq e^{-x}$.

4. Use previous parts to show that

$$p(m,n) \leq \exp\left(-\frac{m(m-1)}{2n}\right).$$

(a) There are n values available for $f(x)$ for every $x \in A$. Hence the total number of function is $|F| = n \cdots n = n^m$.

(b) Let I denote the set of functions that are one-to-one. Let us count $|I|$. Assume that $A = \{x_1, \ldots, x_m\}$. There are m options for $f(x_1)$. Having chosen $f(x_1)$, there are $m-1$ available values for $f(x_2)$. In then same fashion, having chosen $f(x_1), \ldots, f(x_j)$ there are $n-j$ options for $f(x_{j+1})$. From this follows, that

$$|I| = m(m-1)\cdots(m-n+1).$$

This implies that

$$p(m,n) = \mathbb{P}[I] = \frac{m(m-1)\cdots(m-n+1)}{n^m} = \prod_{j=1}^{m-1}\left(1 - \frac{j}{n}\right).$$

Also, if $m > n$, then there is no one-to-one function and hence $p(n,m) = 0$.

(b) Consider the function $f(x) = e^{-x} - 1 + x$. Then $f(0) = 0$ and for $x \geq 0$, we have

$$f'(x) = -e^{-x} + 1 \geq 0.$$

This implies that f is a non-decreasing function and hence $f(x) \geq f(0) = 0$ for $x \geq 0$. This implies that $e^{-x} \geq 1 - x$.

(c) Using part (a) and part (b), for $m \leq n$, we have

$$\begin{aligned} p(m,n) &= \prod_{j=1}^{m-1}\left(1 - \frac{j}{n}\right) = \prod_{j=1}^{m-1}\exp(-\frac{j}{n}) \\ &= \exp\left(-\sum_{j=1}^{m-1}\frac{j}{n}\right) = \exp(\left(-\frac{m(m-1)}{2n}\right). \end{aligned}$$

Tip (Stirling Formula)**.** For large values of n, one can use the following asymptotic formula to approximate $n!$:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

Or equivalently,

$$\log n! \sim \frac{1}{2}\log(2\pi) + (n + \frac{1}{2})\log n - n.$$

Example 1.6.9. 3 couples are sitting on a bench. What is the probability that everyone is sitting next to their partner?

In this case it is clear that $|\Omega| = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 6!$, and $|A| = 6 \cdot 1 \cdot 4 \cdot 1 \cdot 2 \cdot 1$. So

$$\mathbb{P}[A] = \frac{1}{15}.$$

1.7 The Inclusion and Exclusion Principle

Quite often an event can be described as the union or intersection of a number of "simpler" events. In such situations, it would be desirable to have a formula that expressed the probability of the event in question in terms of the probabilities of the simpler events. We will start by some special cases and gradually move to the main theorem of the section.

Theorem 1.7.1. *Let $A_1, \ldots, A_n$ be n mutually disjoint events, i.e. $A_i \cap A_j = \emptyset$, when $i \neq j$. Then*

$$\mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} \mathbb{P}\left[A_i\right].$$

The theorem needs needs to be modified when A_i are not disjoint. However, we have the following inequality, which is useful in bounding the probability from above:

Theorem 1.7.2. *Let $A_1, \ldots, A_n$ be n arbitrary events. Then, we have*

$$\mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \mathbb{P}\left[A_i\right].$$

Proof. We will prove this by induction on n. For $n = 1$ this is obvious and for $n = 2$ follows from the equation $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$. Assume that the inequality is true for n. Given $n+1$ events $A_1, \ldots, A_{n+1}$, we will combine the inequality for two and n events to obtain

$$\begin{aligned}
\mathbb{P}\left[\bigcup_{i=1}^{n+1} A_i\right] &= \mathbb{P}\left[\bigcup_{i=1}^{n} A_i \cup A_{n+1}\right] \\
&\leq \mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] + \mathbb{P}\left[A_{n+1}\right] \qquad (1.1) \\
&\leq \sum_{j=1}^{n} \mathbb{P}\left[A_j\right] + \mathbb{P}\left[A_{n+1}\right] = \sum_{j=1}^{n+1} \mathbb{P}\left[A_j\right].
\end{aligned}$$

□

Inclusion-Exclusion principle

The following theorem is useful for computing the probabilities of events that are described as the union of a finite number of simpler events.

Theorem 1.7.3 (Inclusion-Exclusion principle)**.** *Let $A_1, A_2, \ldots A_n$ be event and $A = \bigcup_{i=1}^{n} A_i$. Then*

$$\mathbb{P}[A] = \sum_{i} \mathbb{P}[A_i] - \sum_{i<j} \mathbb{P}\left[A_i \cap A_j\right] + \sum_{i<j<k} \mathbb{P}\left[A_i \cap A_j \cap A_k\right] - \cdots + (-1)^{n+1} \mathbb{P}\left[A_1 \cap A_2 \cap \cdots \cap A_n\right].$$

Proof. (optional) It is not difficult to give a proof based on the mathematical induction. We will give a short proof which holds when Ω is finite and $\mathbb{P}[A] = \frac{|A|}{\Omega}$. This proof can be modified to hold in general situation. This is left to the interested reader.

Since $\mathbb{P}[A] = \frac{|A|}{|\Omega|}$, it is enough to show

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = \sum_{i=1}^{n} |A_i| - \sum_{i<j} |A_i \cap A_j| + \cdots + (-1)^{n+1} |A_1 \cap A_2 \cap ... \cap A_n|.$$

For each $A \subseteq \Omega$ we define the characteristic function of A which is denoted by $\mathbb{I}_A$ by

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

It is easy to see that:

$$\mathbb{I}_{A \cap B} = \mathbb{I}_A + \mathbb{I}_B,$$

$$\mathbb{I}_{A \cup B} = \mathbb{I}_A + \mathbb{I}_B - \mathbb{I}_A \mathbb{I}_B.$$

Now

$$\mathbb{I}_{A_1 \cup .. \cup A_n} = 1 - (1 - \mathbb{I}_{A_1})(1 - \mathbb{I}_{A_2})...(1 - \mathbb{I}_{A_n})$$

since $x \in A_1 \cup ... \cup A_n$ if $\mathbb{I}_{A_i}(x) = 1$ for some i. This implies that

$$\mathbb{I}_{A_1 \cup ... \cup A_n} = 1 - 1 + \sum_{i=1}^{n} \mathbb{I}_{A_i} - \sum_{i<j} \mathbb{I}_{A_i} \mathbb{I}_{A_j} + ... + (-1)^{n+1} \mathbb{I}_{A_1} ... \mathbb{I}_{A_n} =$$

$$= \sum_{i=1}^{n} \mathbb{I}_{A_i} - \sum_{i<j} \mathbb{I}_{A_i \cap A_j} + ... + (-1)^{n+1} \mathbb{I}_{A_1 \cap ... \cap A_n}$$

Now we have: $|A| = \sum_{x \in \Omega} \mathbb{I}_A(x)$. By summing up the last equality over all the elements of Ω we have

$$\begin{aligned} |A_1 \cup A_2 \cup .... \cup A_n| &= \sum_{x \in \Omega} \mathbb{I}_{A_1 \cup ... \cup A_n}(x) \\ &= \sum_{x \in \Omega} \Big(\sum_{i=1}^{n} \mathbb{I}_{A_i}(x) - \sum_{i<j} \mathbb{I}_{A_i \cap A_j}(x) + ... + (-1)^{n+1} \mathbb{I}_{A_1 \cap \cdots \cap A_n}(x) \Big) \\ &= \sum_{i=1}^{n} |A_i| - \sum_{i<j} |A_i \cap A_j| + \cdots + (-1)^{n+1} \, |A_1 \cap \cdots \cap A_n| \end{aligned}$$

□

Example 1.7.1 (Derangement)**.** n different letters are randomly places in n envelopes with different addresses. What is the probability that no letter is places in the right envelope?

It will be easier to compute the probability of the complement event. Let A_i be the event that the letter i is placed in the right envelope. The event of interest is clearly the complement of

$A = A_1 \cup \cdots \cup A_n$. First note that for each $1 \leq i \leq n$, we have $\mathbb{P}(A_i) = 1/n$. Similarly, for $i_1 < \cdots < i_k$ we have

$$\mathbb{P}\left[A_{i_1} \cap \cdots \cap A_{i_k}\right] = \frac{(n-k)!}{n!} = \frac{1}{n(n-1)\cdots(n-k+1)},$$

because k letters are prescribed to go into k envelopes, but the rest of the letters can go into any envelope. We can now use Theorem 1.7.3 to compute this probability:

$$\begin{aligned}\mathbb{P}[A] &= n \cdot \frac{1}{n} - \binom{n}{2}\frac{1}{n(n-1)} + \cdots + (-1)^{k+1}\binom{n}{k}\frac{(n-k)!}{n!} + \cdots + (-1)^{n+1}\frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1}\frac{1}{n!} = \sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k!}.\end{aligned}$$

From here, we can compute

$$\mathbb{P}[A^c] = \sum_{k=0}^{n}\frac{(-1)^k}{k!}.$$

When n is large, an approximation by an infinite series yield the following neat result:

$$\mathbb{P}[A^c] \approx \sum_{k=0}^{\infty}\frac{(-1)^k}{k!} = \frac{1}{e}.$$

1.8 Geometric Probability

Suppose that a number x is chosen randomly from the interval $[a, b]$. The sample space of this experiment is obviously $\Omega = [a, b]$. How can we associate a probability to an event $A \subseteq \Omega = [a, b]$? It is clear that counting the number of points in the sample space is not a useful approach, as the sample space and most interesting events consist of an infinite number of points.

As the point is chosen from a line segment, natural way of defining probability is to replace the size of a set with the geometric notion of length. In other words, we assume that the probability that the chosen point lands in a given interval $[c, d]$ is given by

$$\mathbb{P}[A] = \frac{\mathcal{L}(A)}{b-a},$$

where $\mathcal{L}$ is the length of A. In order for this to be reasonable definition, we need to check the axioms of probability:

It is indeed easy to see that $\mathbb{P}[A] \geq 0$ and

$$\mathbb{P}[\Omega] = \frac{\mathcal{L}(\Omega)}{b-a} = \frac{b-a}{b-a} = 1.$$

Also, if $A \cap B = \varnothing$ then

$$\mathcal{L}(A \cup B) = \mathcal{L}(A) + \mathcal{L}(B).$$

So

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B].$$

The definition seems to be working. In fact for all of the sets A that we know, we can compute $\mathcal{L}(A)$. If $I = [c, d]$ is an interval, then

$$\mathcal{L}(I) = d - c$$

and in general we can write A as a union of intervals and compute $\mathbb{P}[A]$.

An analogous approach works in higher dimension. The difference is that in higher dimensions, the length needs to be substituted with its appropriate generalization. In dimension two and three, the analogous concepts go by the names of the area and volume. In dimensions $n \geq 4$, one often talks about the n-dimensional volume.

Example 1.8.1. A real number x is chosen in the interval $[-3, 3]$. What is the probability of the event A that $|x - 1| \leq 1$.

The condition $|x - 1| \leq 1$ corresponds to $-1 \leq x - 1 \leq 1$, or $0 \leq x \leq 2$. Hence we have

$$\mathbb{P}[A] = \frac{2}{6} = \frac{1}{3}.$$

Example 1.8.2. Alex and Anna are meeting between noon and 1 pm. Each of them picks a random time in the time interval to show up, wait for 15 minutes and leave. We also assume that they make their decision independently. What is the probability that they meet?

The sample space can be described by

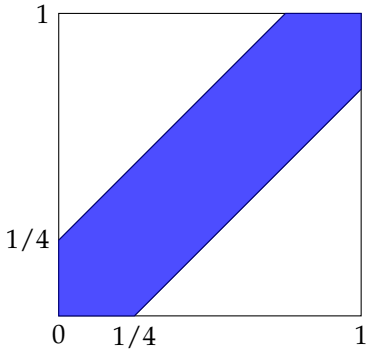
$$\Omega = \{(t_1, t_2) \big| \, 0 \leq t_1 \leq 1, \quad 0 \leq t_2 \leq 1\}$$

where t_1, t_2 are the times that Alex and Anna show up. If M is the event that they meet, then

$$M = \{(t_1, t_2) \big| \, |t_1 - t_2| \leq \frac{1}{4}\}.$$

This can be seen as the shaded area in the square:

$$\mathbb{P}[M] = \text{Area}(M) = 1 - \left(\frac{3}{4}\right)^2 = \frac{7}{16}.$$



Example 1.8.3. A stick of length 1 is folded at 2 random points. What is the probability that one can make a triangle with this stick?

It is easy to see that the sample space is given by

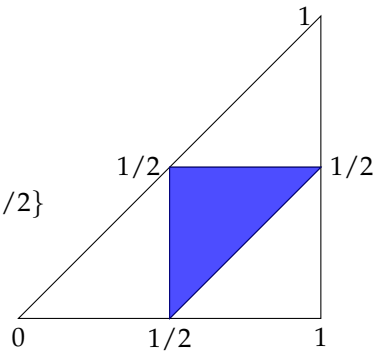
$$\Omega = \{(x,y)\big|\ 0 \le y \le x \le 1\}$$

The set T can then be described by

$$T = \{(x,y)\big|\ x \ge 1/2,\ y \le 1/2,\ x-y \le 1/2\}$$

and hence we have

$$\mathbb{P}[T] = \frac{1}{4}$$



Example 1.8.4 (Bertrand's Paradox)**.** A chord of a circle of radius 1 is chosen randomly. What is the probability of the event E that the length of the cord is at least $\sqrt{3}$?

We will approach this problem in several ways and compute the probability in each case. Interestingly enough, each approach leads to a different answer.

1. Fix one of the endpoints A of the chord. The other endpoint must be in the smaller arc BC for the chord to be longer than $\sqrt{3}$. So $\mathbb{P}[E] = \frac{1}{3}$.
2. Choose a random radius and a point A on the radius and draw a chord whose midpoint is the chosen point A. For the chord to be longer than $\sqrt{3}$, the chord must be closer to the center than the circumference of the circle. So $\mathbb{P}[E] = \frac{1}{2}$.
3. Choose a random point in the circle and draw a chord whose midpoint is the chosen point. Then for the chord to be longer than $\sqrt{3}$ the midpoint must lie outside of a circle of radius $\frac{1}{2}$. So, $\mathbb{P}[E] = \frac{1}{4}$.

Remark 1.8.1. The fact that different methods described above lead to different answers should not come as a surprise. In fact, it would be a surprise (and, perhaps, required some explanation) if one obtained the same answer from some of these methods. Although, each one of the methods promises a way of choosing a "random" chord, the word "random" has different meanings in each method.

To see what is the essence of this "paradox", let us see the following simplified version of Bertrand's paradox. Let, this time, x be a number chosen randomly from the set $[0,1]$. What is the probability p that $x \in [0,1/2]$? The obvious answer is that $p = 1/2$. Yet, let us look at the following outlandish way of choosing a random number: Choose, first, $y \in [0,1]$ randomly as a random seed then set $x = y^{1/n}$, where $n \ge 0$ is a fixed integer. Now,

$$\mathbb{P}[x \le 1/2] = \mathbb{P}[y \le 1/2^n] = \frac{1}{2^n}.$$

This shows that a liberal usage of the word "random" can easily lead to confusion. In the following chapters, by introducing the notion of random variables, we will see how such confusions can be avoided.

Example 1.8.5 (Buffon's needle)**.** A needle of length 1 is randomly dropped on a plane which is ruled by parallel lines with distance 1 between any two consecutive ones. Compute the probability that it hits one of the lines.

Without loss of generality, let us assume that the needle lands so that the middle point is in the region

$$A = \left\{ (x, y) \middle| \, 0 \leq y \leq \frac{1}{2} \right\}.$$

Then the needle's position can be determined by knowing the angle θ that it forms with the x-axis. Clearly $0 \leq \theta \leq \pi$. Now let's see when the needle is going to hit the line.

The condition is that

$$y \leq \frac{1}{2} \sin \theta,$$

where

$$\Omega = \{ (y, \theta) \big| \, 0 \leq y \leq \frac{1}{2}, \, 0 \leq \theta \leq \pi \}.$$

Note that Ω is a rectangle of area $\pi/2$. Now, let H denote the event that the needle hits one of the lines. By the above computation, we have:

$$\mathbb{P}[H] = \frac{1}{\text{Area}(\Omega)} \int_0^\pi \frac{1}{2} \sin \theta d\theta = \frac{2}{\pi}.$$

Remark 1.8.2. A similar computation shows that if the length of the needle is $l \geq 1$, the probability of hitting one of the lines is given by $2l/\pi$. In particular, for $l = \pi/4$, the hitting probability is $1/2$. Such a needle can hence be used as a surrogate for the fair coin.

2
Conditional Probability and Independence

> If the Probability that an Event shall Happen be $\frac{1}{r}$, and if that Event being supposed to have Happened, the Probability of another Happening be $\frac{1}{s}$; the Probability of both happening will be $\frac{1}{r} \times \frac{1}{s}$ or $\frac{1}{rs}$.
>
> The Doctrine of Chance,
> Abraham de Moivre

2.1 Conditional Probability

In the previous chapter, we became familiar with the concept of probability as way of quantifying randomness. It is a function that associates to every subset A of a sample space Ω a number $P(A)$ between 0 and 1, which expresses the likelihood of A. Now, suppose we have two events A and B and we would like to understand the relation between the occurrence of A and that of B. Intuitively, $\mathbb{P}[A]$ is all the information about the likelihood of A happening. Now, suppose that we are told that the event B has already taken place. It is obvious that using this extra piece of information we can update our guess about the likelihood of A. (For two extreme cases, think of $B = A$ or $B = A^c$. Now the updated probabilities are 1 and 0, regardless of what $\mathbb{P}[A]$ was initially was. We will formalize this concept in the following definition:

Definition 2.1.1. Suppose that A, B are two events and that $\mathbb{P}(B) \neq 0$. The conditional probability $\mathbb{P}(A|B)$ (read as A given B) is defined by

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Remark 2.1.2. Note that for this definition to make sense, one needs to assume that $\mathbb{P}[B] \neq 0$.

Example 2.1.1. Two fair dice are rolled. If the sum of the resulted numbers is 7, what is the probability that the smaller number is at least 3.

It is easy to see that the sample space is given by

$$\Omega = \{(i,j) : 1 \leq i,j \leq 6\}.$$

Let A denote the event that $i+j=7$ and B the event that $\min(i,j) \geq 3$. Clearly, we have

$$A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}, \quad A \cap B = \{(3,4),(4,3)\}.$$

From here we have

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = \frac{2/36}{6/36} = \frac{1}{3}.$$

Remark 2.1.3. In the special case of uniform probability, one does not need to compute the size of the sample space. In fact, assuming that $\mathbb{P}[A] = |A|/|\Omega|$ for all $A \subseteq \Omega$, we have

$$\mathbb{P}[A|B] = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} = \frac{|A \cap B|}{|B|}.$$

This can be useful in situations in which computing $|\Omega|$ is not easy.

2.2 *Conditioning*

One of the applications of conditional probability is *conditioning* which is an effective method for computing certain probabilities.

Suppose we want to compute $\mathbb{P}[A]$ for some $A \subseteq \Omega$. In the case that Ω is a finite set, this is essentially a counting problem. In many situations, it may be very difficult to compute $|A|$ directly, but one can cut A into a number of pieces whose size is easier to compute. A mathematical description of this situation can be given by a partition $\Omega = B_1 \cup B_2 \cdots \cup B_n$ of the sample space.

So we can write:

$$\begin{aligned}\mathbb{P}[A] &= \mathbb{P}[\cup_{i=1}^{n}(A \cap B_i)] \\ &= \sum_{i=1}^{n} \mathbb{P}[A \cap B_i] \\ &= \sum_{i=1}^{n} \mathbb{P}[A|B_i]\,\mathbb{P}[B_i].\end{aligned}$$

So we obtain the following useful formula:

Theorem 2.2.1 (Conditioning). *Let $\Omega = B_1 \cup B_2 \cdots \cup B_n$ be a partitioning of the sample space and A be an event. Then*

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A|B_i]\,\mathbb{P}[B_i].$$

This will become more clear with an example.

Example 2.2.1. Alex has 5 coins in his pocket.Two are double-headed. one is double-tailed and the other two are normal.One of the coins is randomly chosen and flipped.

1. What is the probability that the outcome is heads?

2. He opens his eyes and sees that the outcome is heads.What is the probability that the flipped coin is double-headed?

Let B_{HH}, B_{TT}, B_{HT} denote the events that the selected coin is double-headed, double-tailed or normal.

Let A be the event that the outcome is heads

$$\begin{aligned}\mathbb{P}[A] &= \mathbb{P}[A|B_{HH}]\,\mathbb{P}[B_{HH}] + \mathbb{P}[A|B_{TT}]\,\mathbb{P}[B_{TT}] + \mathbb{P}[A|B_{HT}]\,\mathbb{P}[B_{HT}] \\ &= 1\frac{2}{5} + 0 + \frac{1}{2}\frac{2}{5} = \frac{3}{5} \\ \mathbb{P}[B_{HH}|A] &= \frac{\mathbb{P}[A|B_{HH}]\,\mathbb{P}[B_{HH}]}{\mathbb{P}(A)} \\ &= \frac{\frac{2}{5}}{\frac{3}{5}} = \frac{2}{3}\end{aligned}$$

Example 2.2.2. Two subsets S, T of the set $X = \{1, 2, \ldots, n\}$ are randomly and independently chosen. Compute the probability that $S \subseteq T$.

The sample space is clearly given by

$$\Omega = \{(S, T) : S, T \subseteq X\}.$$

So $|\Omega| = 2^n \cdot 2^n = 4^n$. The desired event can be described by

$$A = \{(S, T) : S \subseteq T \subseteq X\}.$$

Let $B_i = \{(S, T)||T| = i\}$ for $i = 0, 1, .., n$. Clearly, B_i partition the sample space. Using Theorem 2.2.1, we have

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A|B_i]\,\mathbb{P}[B_i] = \sum_{i=1}^{n} \frac{2^i}{2^n}\frac{\binom{n}{i}}{2^n} = \frac{1}{4}\sum_{i=0}^{n}\binom{n}{i}2^i = \frac{1}{4^n}(2+1)^n = \frac{3^n}{4^n}.$$

Later, we will see a different way of computing this probability.

Example 2.2.3. An urn contains r red and b blue balls. A ball is drawn from the urn and discarded.

1. What is the probability that the discarded ball is blue?

2. Without knowing the color of the first color, what is the probability that a second ball drawn is blue?

Let R_j and B_j denote the events that the jth ball are red and blue, respectively. Then

$$\mathbb{P}[B_2] = \mathbb{P}[B_2|B_1]\,\mathbb{P}[B_1] + \mathbb{P}[B_2|R_1]\,\mathbb{P}[R_1]$$

Since we have

$$\mathbb{P}[B_1] = \frac{b}{b+r}, \mathbb{P}[R_1] = \frac{r}{b+r}$$

and

$$\mathbb{P}[B_2|B_1] = \frac{b-1}{b+r-1}, \quad \mathbb{P}[B_2|R_1] = \frac{b}{b+r-1}$$

By substituting in the formula above, we obtain

$$\mathbb{P}[B_2] = \frac{b}{b+r}.$$

An interesting remark is that in the absence of information about the color of the first ball, the odds that the second ball is blue is the same as the first ball being blue.

Example 2.2.4. Alice and Bob play the game of "heads and tails". At every round of the game, a coin is flipped. If it lands heads, Alice ones 1 Euro form Bob and if it lands tails, Bob wins 1 Euro from Alice. They continue playing the game until one of them reaches zero. Suppose that the initial wealth of Alice and Bob are a and b Euros, respectively. What is the probability that Alice/Bob wins the game.

In order to solve the problem, we will denote by A_i(B_i, respectively) the events that Alice (Bob, respectively) will lose if she starts the game with i euros. We will also set $p_i = \mathbb{P}[A_i]$. We will use conditioning on the outcome of the first game. Let F denote the event that Alice loses the first game. Then, assuming that $i \geq 1$, we can write

$$\mathbb{P}[A_i] = \mathbb{P}[A_i|F]\,\mathbb{P}[F] + \mathbb{P}[A_i|F^c]\,\mathbb{P}[F^c] = \frac{1}{2}(p_{i-1} + p_{i+1}).$$

This shows that

$$p_{i+1} - p_i = p_i - p_{i-1}.$$

Calling the common value of these differences c, we have $p_i = p_0 + ci = 1 + ci$. On the other hand, from $p_0 = 1$ and $p_{a+b} = 0$, we can deduce that

$$p_i = \frac{a+b-i}{a+b}.$$

In particular, we have

$$p = \frac{a}{a+b}.$$

An analogous argument also shows that the probability that Bob wins the game is $q = b/a + b$. Since $p + q = 1$, we can also see that the game ends with probability one.

2.3 *Bayes' Formula*

Imagine a real-world situation that an event A can be caused by different events $B_1, \ldots, B_n$. We would like to compute the probability of the event B_i, in light of the evidence that A has occurred.

First, to show the idea, assume $\Omega = B_1 \cup B_2$ is a partition of Ω. We would like to compute $\mathbb{P}[B_1|A]$.

We have:

$$\mathbb{P}[B_1|A] = \frac{\mathbb{P}[B_1 \cap A]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A|B_1]\,\mathbb{P}[B_1]}{\mathbb{P}[A|B_1]\,\mathbb{P}[B_1] + \mathbb{P}[A|B_2]\,\mathbb{P}[B_2]}.$$

This is a special case of what is called the **Bayes' formula**.

More generally, we have the following theorem:

Theorem 2.3.1 (Bayes' Formula). *Let $\Omega = B_1 \cup B_2 \cup \cdots \cup B_n$ be a partitioning of the sample space Ω. Then we have*

$$\mathbb{P}[B_i|A] = \frac{\mathbb{P}[A|B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^{n} \mathbb{P}\left[A|B_j\right]\mathbb{P}\left[B_j\right]}.$$

Example 2.3.1. Through a transmission channel two types of messages can be sent: 0 and 1. We assume that 40% of the time a 1 is transmitted. The probability that 0 is correctly received is 0.80 and the probability that a transmitted 1 is correctly received is 0.90. Determine

a) the probability of a 0 being received.

b) given a 1 received, the probability that 1 was transmitted.

Let T_0 and T_1 be the events that a 0,1 is transmitted and R_0, and R_1 be the events that a 0,1 is received. Then the information given in the problem translates to $\mathbb{P}[T_1] = \frac{4}{10}$ and $\mathbb{P}[T_0] = \frac{6}{10}$ and the following conditional probabilities:

$$\mathbb{P}[R_0|T_0] = \frac{8}{10}, \quad \mathbb{P}[R_1|T_0] = \frac{2}{10}, \quad \mathbb{P}[R_1|T_1] = \frac{9}{10}, \quad \mathbb{P}[R_0|T_1] = \frac{1}{10}.$$

So

$$\mathbb{P}[R_0] = \mathbb{P}[R_0|T_0]\,\mathbb{P}[T_0] + \mathbb{P}[R_0|T_1]\,\mathbb{P}[T_1] = \frac{8}{10}\cdot\frac{6}{10} + \frac{1}{10}\cdot\frac{4}{10} = \frac{52}{100}.$$

$$\mathbb{P}[T_1|R_1] = \frac{\mathbb{P}[R_1|T_1]\,\mathbb{P}[T_1]}{\mathbb{P}[R_1|T_1]\,\mathbb{P}[T_1] + \mathbb{P}[R_1|T_0]\,\mathbb{P}[T_0]} = \frac{\frac{9}{10}\cdot\frac{4}{10}}{\frac{9}{10}\cdot\frac{4}{10} + \frac{2}{10}\cdot\frac{6}{10}} = \frac{36}{48} = \frac{3}{4}.$$

Example 2.3.2 (Fallacy of "Confusion of inverse"). Suppose 5% of all cancers are malignant and suppose we have a test that is 90% accurate in determining malignancy. Suppose further that a test result has come back positive. Find the probability that the tumor is malignant.

Let M and P denote the events that the tumor is malignant and that the test is positive.

We know $\mathbb{P}[P|M] = 0.90$ and $\mathbb{P}[P|M^c] = 0.10$. So

$$\begin{aligned}
\mathbb{P}[M|P] &= \frac{\mathbb{P}[P|M]\,\mathbb{P}[M]}{\mathbb{P}[P|M]\,\mathbb{P}[M] + \mathbb{P}[P|M^c]\,(\mathbb{P}[M^c]} \\
&= \frac{\frac{90}{100}\frac{5}{100}}{\frac{90}{100}\frac{5}{100} + \frac{10}{100}\frac{95}{100}} \\
&= \frac{450}{1400} = 0.32
\end{aligned}$$

2.4 *Independence*

Recall from the previous section that for events A and B, the conditional probability of A given B is defined by

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

As much as $\mathbb{P}[A]$ indicates the degree of belief in occurrence of A, the conditional probability $\mathbb{P}[A|B]$ can be viewed as the degree of belief in the occurrence of A, under the assumption that B has taken place. In other words, upon learning the extra information on the occurrence of B, all the other probabilities need to be updated. Now, the updated probability may happen to be the same as the $\mathbb{P}[A|B] = \mathbb{P}[A]$, then the extra information on B has no influence on our degree of belief about A. It is natural to say that A and B are independent in such a situation. Using the definition of conditional probability, this will translate to $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. The notion of independence makes sense for more than two events. We will now give the official definition:

Definition 2.4.1. Let $A_1, \ldots, A_n$ denote events in the sample space Ω. We say that $A_1, \ldots, A_n$ are independent, if for any $k \geq 2$ and any indices $1 \leq i_1 < i_2 < \cdots < i_k \leq n$, we have

$$\mathbb{P}\left[A_{i_1} \cap \cdots \cap A_{i_k}\right] = \mathbb{P}\left[A_{i_1}\right] \cdots \mathbb{P}\left[A_{i_k}\right].$$

Example 2.4.1. Let us unpack this definition to see what it entails for small values of n. For two events A, B, independence is equivalent to $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. For three events A, B, C, one requires the following four equalities:

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B], \quad \mathbb{P}[A \cap C] = \mathbb{P}[A]\mathbb{P}[C], \quad \mathbb{P}[B \cap C] = \mathbb{P}[B]\mathbb{P}[C], \quad \mathbb{P}[A \cap B \cap C] = \mathbb{P}[A]\mathbb{P}[B]\mathbb{P}[C].$$

Example 2.4.2. Assume that a pair of coins is flipped. So

$$\Omega = \{HH, HT, TH, TT\}$$

Suppose that the probabilities of different points in the sample space are given by

$$\mathbb{P}[HH] = \frac{1}{3}$$

$$\mathbb{P}[HT] = \frac{1}{6}$$

$$\mathbb{P}[TH] = \frac{1}{4}$$

$$\mathbb{P}[TT] = \frac{1}{4}$$

Let A and B be the events that the first and second coin land heads, respectively. Then

$$\mathbb{P}[A] = \mathbb{P}[\{HH, HT\}] = \frac{1}{2}$$

$$\mathbb{P}[B] = \mathbb{P}[\{HH, TH\}] = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$$

$$\mathbb{P}[A \cap B] = \mathbb{P}[\{HH\}] = \frac{1}{3},$$

since $\frac{1}{3} \neq \frac{1}{2} \cdot \frac{7}{12}$. So A and B are not independent.

Example 2.4.3. Note that the pairwise independence does not imply independence. To see this, consider two independent fair coins. The sample space corresponding to this experiment is obviously given by

$$\Omega = \{HH, HT, TH, TT\}.$$

Consider the following events: For $i = 1, 2$, let A_i be the event defined by the condition that the i-th throw is H. Hence

$$A_1 = \{HH, HT\}, \quad A_2 = \{TH, HH\}.$$

Also, let A_0 denote the event that there are an event number of heads, hence

$$A_0 = \{HH, TT\}.$$

It is easy to see that A_0, A_1, A_2 are pairwise independent, but they are not independent as

$$\mathbb{P}[A_0 \cap A_1 \cap A_2] = \frac{1}{4} \neq \mathbb{P}[A_0]\,\mathbb{P}[A_1]\,\mathbb{P}[A_2].$$

Example 2.4.4. An unfair coin lands heads with probability $0 < p < 1$ and tails with probability $1 - p$. The coin is flipped twice. Assuming that the outcome of the first and the second throw are independent, we have

$$\mathbb{P}[HT] = \mathbb{P}[H* \cap *T] = \mathbb{P}[H*]\,\mathbb{P}[*T] = p(1-p).$$

Similarly, one can also see that $\mathbb{P}[TH] = p(1-p) = \mathbb{P}[HT]$. This equality allows one to use an unfair coin in designing a fair way of choosing between alternatives. The coin is thrown twice. The first part wins if the outcomes is HT and the second party wins if the outcome is TH. In other cases, the process needs to be repeated. Note that the probability of the event R_1 that the process gets repeated is

$$\mathbb{P}[R] = \mathbb{P}[\{HH, TT\}] = p^2 + (1-p)^2.$$

Set $q = p^2 + (1-p)^2 < 1$. Similarly, the probability that the second round does not result in a winner is $\mathbb{P}[R_2] = q$. Again, using independence, we have

$$\mathbb{P}[R_1 \cap R_2] = q^2.$$

A similar argument shows that

$$\mathbb{P}[R_1 \cap \cdots \cap R_n] = q^n.$$

Let R denote the event that the process remains inconclusive forever. It is clear that $R \subseteq R_1 \cap \cdots \cap R_n$ for every $n \geq 1$. Hence, $\mathbb{P}[R]\, q^n$, for all n, which implies that $\mathbb{P}[R] = 0$. In other words, with probability 1, the process will eventually end.

Example 2.4.5. A number n is randomly and uniformly chosen from the set $X = \{1, 2, 3, 4, ...30\}$. Let A_2, A_3, A_7 denote the events that n is divisible by $2, 3, 7$, respectively. Prove that A_2, A_3 are independent, but neither A_2 and A_{19} nor A_3 and A_{19} are independent.

It is clear that $|A_2| = 15$ and $|A_3| = 10$. Hence

$$\mathbb{P}[A_2] = \frac{1}{2}, \qquad \mathbb{P}[A_3] = \frac{1}{3}$$

On the other hand

$$\mathbb{P}[A_{19}] = \frac{1}{30}.$$

Now we can compute:

$$\mathbb{P}[A_2 \cap A_3] = \frac{|\{6, 12, 18, 24, 30\}|}{30} = \frac{1}{6} = \mathbb{P}[A_2]\,\mathbb{P}[A_3]$$

$$\mathbb{P}[A_2 \cap A_{19}] = 0 \neq \mathbb{P}[A_2]\,\mathbb{P}[A_{19}]$$

$$\mathbb{P}[A_3 \cap A_{19}] = 0 \neq \mathbb{P}[A_3]\,\mathbb{P}[A_{19}].$$

Even though it seems that divisibility by 2 and divisibility by 19 are not independent, it has to be said that this is due to considering only the integers up to 30. Suppose p and q are distinct prime numbers and let us also replace the sample space by $\Omega = \{1, 2 \ldots, n\}$, then

$$\mathbb{P}[A_p] = \frac{1}{n}\lfloor\frac{n}{p}\rfloor \approx \frac{1}{p}, \qquad \mathbb{P}[A_q] = \frac{1}{n}\lfloor\frac{n}{q}\rfloor \approx \frac{1}{q}.$$

Moreover, a number is divisible by both p and q iff it is divisible by pq. This shows that

$$\mathbb{P}[A_p \cap A_q] = \frac{1}{n}\lfloor\frac{n}{pq}\rfloor \approx \frac{1}{pq}.$$

This shows that the A_p and A_q will asymptotically be independent as $n \to \infty$.

Example 2.4.6. Anja tosses a fair coin $n + 1$ times. Alex tosses the same coin n times. What is the probability that Anja gets more heads than Alex.

We will give two solutions. Let H_1 and H_2 denote the number of heads obtained by Anja and Alex. Similarly, T_1 and T_2 denote the number of tails obtained by Anja and Alex. Since the coin is fair, we have

$$\mathbb{P}[H_1 > H_2] = \mathbb{P}[T_1 > T_2].$$

On the other hand

$$\mathbb{P}[T_1 > T_2] = \mathbb{P}[n + 1 - H_1 > n - H_2] = \mathbb{P}[H_1 \leq H_2].$$

This implies that

$$\mathbb{P}[H_1 > H_2] = \mathbb{P}[H_1 \leq H_2].$$

Since $\mathbb{P}[H_1 > H_2] + \mathbb{P}[H_1 \leq H_2] = 1$, we have

$$\mathbb{P}[H_1 > H_2] = \frac{1}{2}.$$

3
Random variables

3.1 What is a random variable?

When we use probability theorem in real life, our sample space can be viewed as the set of all possible scenarios. Consider, for instance, the common question of modeling the stock market. Each point ω in the sample space can be viewed as a possible state of the world at some point in the future. As this example shows, we are typically not interested in ω itself, but rather in quantities that depend on ω. A typical example is the price of a stock S, which depends on the state of the world ω, and hence can be viewed as a function on the sample space Ω. More generally, we are interested in assigning a numerical quantity to an outcome $\omega \in \Omega$ of the experiment that captures one particular aspect. This leads to the following definition.

Definition 3.1.1. Let $(\Omega, \mathbb{P})$ be a probability space. A function

$$X : \Omega \to \mathbb{R}$$

is called a real valued *random variable*. Similarly, a function $X : \Omega \to \mathbb{R}^n$ is called a vector-valued random variable.

Remark 3.1.2. In a more rigorous treatment of probability theory, one needs to impose some regularity condition known as *measurability* on X. This regularity condition is needed for working with random variables. As all the functions we consider in this course are measurable, we will not be our concerned with the issue in this course.

Remark 3.1.3. One can equally consider random variables that take values in the set of complex numbers, arbitrary vector spaces, or even more general spaces. Although the development of the theory will not be significantly different for such random variables, we will not do so now. In the next chapter, we will (implicitly) have to deal with random variables that take values in $\mathbb{R}^n$.

Example 3.1.1. The possible outcome of the rolling of a fair die are $\Omega = \{1, 2, 3, 4, 5, 6\}$. Define X by $X(1) = X(3) = X(5) = 1$ and $X(2) = X(4) = X(6) = 0$. This defines a Bernoulli random variable with $\mathbb{P}[X = 1] = \mathbb{P}[X = 0] = 1/2$.

Example 3.1.2. Suppose that the flipping of a coin can result in heads with probability p and in tails with probability $1 - p$, This coin is tossed n times and we further assume that the outcomes

are independent. So here

$$\Omega = \{(x_1, x_2, \ldots, x_n) \mid x_i = 0 \text{ or } 1\}.$$

Each outcome of this experiment corresponds to an element $\omega \in \Omega$. Let

$$\begin{aligned} X_1(\omega) &= \{\text{first head}\}, \\ X_2(\omega) &= \{\text{first tail}\}, \\ X_3(\omega) &= \{\text{total number of heads}\}, \\ X_4(\omega) &= \{\text{total number of tails}\}. \end{aligned}$$

Even though, the definition of a random variable makes a reference to a probability space, we will often makes no reference to the underlying probability space. In fact, the interesting properties of random variables that those that are independent of the probability space on which it is defined and can be expressed in terms of its distribution function. This is the reason that in the future we will only talk about concepts such as cumulative distribution function, probability density function, or probability mass function of random variables, without mentioning Ω which is pushed to the background. In all cases, one can easily construct such an Ω, though there is often no canonical choice for such a space.

3.2 *Discrete Random Variables*

In this section, we will discuss several discrete random variables.

Definition 3.2.1. The probability mass function of a random variable X with sets of values $x_1, \ldots$ is defined by

$$p(x) = \mathbb{P}[X = x].$$

Note that for every i, we have $p(x_i) = p_i > 0$.

Remark 3.2.2. Note that the probability mass function must satisfy the following condition:

$$\sum_{j=1}^{\infty} p(x_j) = 1.$$

Example 3.2.1. (Bernoulli variable) One of the most notable random variables is a random variable X that takes two value. It is common to take these two values to be zero and one.

Definition 3.2.3. A random variable X is called the *Bernoulli* random variable with parameter p if it only takes values 0 and 1, and

$$\mathbb{P}[X = 1] = p, \qquad \mathbb{P}[X = 0] = 1 - p.$$

One usually use Bernoulli random variables to model the events that have exactly two outcome, e.g. Heads/Tails, success/failure, upmarket move/downmarket move, etc. For instance,

take $\Omega = \{H, T\}$, representing the possible outcomes of a coin flip. Assuming that $\mathbb{P}[\{H\}] = p$, we have $\mathbb{P}[\{T\}] = 1 - p$. Now, let X be defined by $X(H) = 1$ and $X(T) = 0$. Clearly X is a Bernoulli random variable.

Example 3.2.2. (Binomial Distribution) Let Ω be the set of all sequences of length n of the letters H and T, standing for Heads and Tails. Clearly $|\Omega| = 2^n$. Each element of Ω can be seen as the outcome of n flips a coin. For every element $\omega \in \Omega$, let $X(\omega)$ be the number of appearances of H in ω. Define a probability measure on Ω by assigning

$$\mathbb{P}[\{\omega\}] = p^{X(\omega)}(1-p)^{n-X(\omega)}.$$

It is easy to verify that this defines a probability measure. Consider the map $X : \Omega \to \mathbb{R}$. Clearly X takes only values $0, 1, \ldots, n$ and for $0 \leq k \leq n$ we have

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

A random variable that satisfies the above equation is called a **Binomial** random variable with parameter (n, p).

Definition 3.2.4. A random variable X has the Binomial distribution with parameters (n, p) if,

$$\mathbb{P}[X = k] = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } 0 \leq k \leq n \\ 0 & \text{otherwise} \end{cases}$$

Example 3.2.3 (Geometric Distribution)**.** Consider a coin that turns up heads with probability p and tails with probability $1 - p$. The coin is flipped until the a heads shows ups. Let X be the number of the flips needed. Assuming that the outcome of the flips are independent, we have

$$\mathbb{P}[X = n] = (1-p)^{n-1} p.$$

Definition 3.2.5. A discrete random variable X has Geometric distribution with parameter p if

$$\mathbb{P}[X = k] = \begin{cases} p^k (1-p) & \text{if } k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Example 3.2.4 (Poisson distribution)**.** Suppose that particles can be created at random time. Also suppose that the probability that a particle is created in a time segment $[t_0, t_1]$ depends only on $t_1 - t_0$, not on t_0 and not on how many particles are already created. Also, assume that the probability of a particle being created in time Δt is approximately equal to $\lambda \cdot \Delta t$, for small Δt. Such process of particle creation is called the *Poisson process*. Many random processes happening in real life are modeled by Poisson processes. Let X be the number of particles created in time t. Then we have

$$\mathbb{P}[X = k] = \lim_{n \to \infty} \binom{n}{k} (\lambda t/n)^k (1 - \lambda t/n)^{n-k} = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

This distribution is called the Poisson distribution.

Definition 3.2.6. A random variable X has the Poisson distribution with parameter λ if

$$\mathbb{P}[X = k] = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda}. & \text{if } k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

3.3 *Distribution functions*

One of the important functions that can be associated to a random variable is the distribution function.

Definition 3.3.1. Let $X : \Omega \to \mathbb{R}$ be a random variable. The *probability distribution function* of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(t) = \mathbb{P}[X \leq t].$$

The following theorem summarizes the most basic properties of distribution functions.

Theorem 3.3.2. *The probability distribution function enjoys the following properties:*

1. *F_X is (non-strictly) increasing: if $t_1 \leq t_2$, then $F_X(t_1) \leq F_X(t_2)$.*
2. *F_X is right-continuous, that is, for every $t \in \mathbb{R}$:*

$$\lim_{s \to t+} F_X(s) = F_X(t).$$

3. *F_X has limits at $\pm\infty$, namely, $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.*

Remark 3.3.3. Note that the distribution function allows one to compute the probabilities of the events of the form $X \leq t$. From here one can immediately see that

$$\mathbb{P}[X > t] = 1 - \mathbb{P}[X \leq t] = 1 - F_X(t).$$

On the other hand, computing $\mathbb{P}[X < t]$ requires more work. We will first need a notation. If $f(t)$ is a function of t, the left-limit $\lim_{s \to t-} f(s)$, when exists, will be denoted by $f(s-)$.

Proposition 3.3.4. *For a random variables X and $t \in \mathbb{R}$, we have*

$$\mathbb{P}[X < t] = F_X(t-) := \lim_{s \to t-} F_X(s).$$

Proof. Let A_n be the event defined by $X \leq t - \frac{1}{n}$, and A the event $X < t$. It is easy to see that $A = \bigcup_{j=1}^{\infty} A_j$. Since $A_1 \subseteq A_2 \subseteq \cdots$, we have

$$\mathbb{P}[A] = \lim_{n\to\infty} \mathbb{P}[A_n] = \lim_{n\to\infty} F_X(t - \frac{1}{n}) = \lim_{s\to t-} F_X(s).$$

□

Corollary 3.3.5. *For a random variables X and $s, t \in \mathbb{R}$, we have*

1. $\mathbb{P}[X \geq t] = 1 - F_X(t-)$.
2. $\mathbb{P}[X = t] = F_X(t) - F_X(t-)$.

Example 3.3.1. Let X be a random variable with distribution function F_X. Find the probabilities of the following events:

$$a < X \leq b, \quad a < X < b, \quad X = a.$$

in terms of values and/or limit values of F_X.

Using the definition of the distribution function and Corollary 3.3.5, we can write

$$\mathbb{P}[a < X < b] = \mathbb{P}[X < b] - \mathbb{P}[X \leq a] = F_X(b-) - F_X(a).$$

$$\mathbb{P}[a < X \leq b] = \mathbb{P}[X \leq b] - \mathbb{P}[X \leq a] = F_X(b) - F_X(a).$$

$$\mathbb{P}[X = a] = \mathbb{P}[X \leq a] - \mathbb{P}[X < a] = F_X(a) - F_X(a-).$$

Remark 3.3.6. Note that for a random variable X, it can happen that $\mathbb{P}[X = c] > 0$ for some c. Such points correspond to the discontinuities of the distribution function F_X. On the other hand, assuming that F_X is a continuous function, we have $\mathbb{P}[X = c] = 0$, for every $c \in \mathbb{R}$.

3.4 Continuous Random Variables

In this section we will introduce the notion of continuous random variables and provide a number of useful examples.

Definition 3.4.1. A random variable $X : \Omega \to \mathbb{R}$ is called *continuous* if there exists a non-negative function $f_X : \mathbb{R} \to \mathbb{R}$, called the *probability density function* of X, such that for all $t \in \mathbb{R}$, we have

$$F_X(t) = \int_{-\infty}^{t} f_X(x)dx.$$

In fact more is true. Let A be a subset of $\mathbb{R}$ consisting as a union of intervals. Then

$$\mathbb{P}[X \in A] = \int_A f_X(x)dx.$$

The definition can be viewed as the special case $A = (-\infty, t]$. The Fundamental Theorem of Calculus proves the relation between the probability density function and distribution function of a continuous random variable.

Theorem 3.4.2. *If F_X is differentiable, then*

$$f_X(t) = \frac{d}{dt} F_X(t).$$

Recall that the probability density function of a continuous random variable satisfies the following properties:

D1 $f_X \geq 0$.

D2 $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Any function $f : \mathbb{R} \to \mathbb{R}$ satisfying D1 and D2 can serve as the density function of a continuous random variables. We will now consider a number of density functions that one encounters quite often in practice.

Example 3.4.1 (Uniform distribution)**.** We say that a random variable X is uniformly distributed in the interval $I = [a, b]$ if the probability that X belongs to a segment $I \subseteq [a, b]$ is proportional to the length of I. This leads to the following definition:

Definition 3.4.3. A random variable X has uniform distribution over the interval $[a, b]$, if its probability density function is given by

$$f_X(t) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq t \leq b \\ 0 & \text{otherwise} \end{cases}$$

The distribution function of X can be computed by a simple integration.

Proposition 3.4.4. *The distribution function of a random variable X with uniform distribution over the interval $[a, b]$ is given by*

$$F_X(t) = \begin{cases} 0 & \text{if } t \leq a \\ \frac{t-a}{b-a} & \text{if } a \leq t \leq b \\ 1 & t \geq b. \end{cases}$$

Proof. Since $f_X(t) = 0$ for $t \notin [a, b]$, we clearly have $F_X(t) = 0$ for $t \leq a$ and $F_X(t) = 1$ for $t \geq b$. For $a \leq t \leq b$, we have

$$F_X(t) = \int_{-\infty}^{t} f_X(s)ds = \int_a^t \frac{1}{b-a} ds = \frac{t-a}{b-a}.$$

□

Example 3.4.2. Suppose θ has uniform distribution over the interval $[-\pi/2, \pi/2]$ and $X = \sin\theta$. Compute the probability density function of X.

First note that for $-\pi/2 \leq t \leq \pi/2$, we have

$$\mathbb{P}[\theta \leq t] = \frac{2t - \pi}{2\pi}.$$

As the sine function is increasing over the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$, and takes values in $[-1, 1]$, we have

$$F_X(t) = \mathbb{P}[X \leq t] = \mathbb{P}[\sin\theta \leq t] = \mathbb{P}\left[\theta \leq \sin^{-1} t\right] = \frac{2\sin^{-1} t - \pi}{2\pi}.$$

Differentiating with respect to t gives the probability density function of X:

$$f_X(t) = \frac{1}{\pi}\sqrt{1 - t^2}, \quad -1 \leq t \leq 1.$$

Hence, we have

$$f_X(t) = \begin{cases} \frac{1}{\pi}\sqrt{1-t^2} & \text{if } -1 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Example 3.4.3. Let X be uniformly distributed in $[-1, 1]$, and $Y = X^2$. Let us compute the distribution function of Y. Note that in this examples, the map $t \mapsto t^2$ is not one-to-one on the interval $[-1, 1]$. On the other hand, Y takes values in $[0, 1]$. From here, for $0 \leq t \leq 1$, we have

$$F_Y(t) = \mathbb{P}[Y \leq t] = \mathbb{P}\left[X^2 \leq t\right] = \mathbb{P}\left[0 \leq X \leq \sqrt{t}\right] + \mathbb{P}\left[-\sqrt{t} \leq X \leq 0\right] = \sqrt{t}.$$

The corresponding probability density is:

$$f_X(x) = F_X'(x) = \frac{1}{2\sqrt{x}}, \quad 0 \leq x \leq 1.$$

It goes without saying that $f_X(x) = 0$ if x is outside of $[0, 1]$.

Example 3.4.4 (Exponential distribution)**.** The next class of continuous random variables that we will consider is the exponential distribution. We will first give the definition and then prove a property that characterizes this random variable.

Definition 3.4.5. A continuous random variable X has an exponential distribution with parameter λ if

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

We can easily compute the distribution function of such a random variable:

Proposition 3.4.6. *The distribution function of a random variable X with geometric distribution with parameter λ is given by*

$$F_X(t) = \begin{cases} 1 - e^{-\lambda t} & \textit{if } t \geq 0 \\ 0 & \textit{otherwise} \end{cases}.$$

Proof. Obviously, we will only need to consider $t \geq 0$. Then we have,

$$F_X(t) = \int_0^s \lambda e^{-\lambda s} ds = 1 - e^{-\lambda t}.$$

□

Geometric random variables enjoy a particular property that makes them distinguished among other random variables. We will first give a definition.

Definition 3.4.7. Let X be a random variable X taking positive values. We say that X is memory-less if for all $t, s > 0$, we have

$$\mathbb{P}\left[X \geq s + t | X \geq s\right] = \mathbb{P}\left[X \geq t\right].$$

The significance of the exponential distribution is due to the following theorem.

Theorem 3.4.8. *Let X be a non-negative continuous random variable. Then X has an exponential distribution if and only if X is memory-less*

Proof. First let us prove the easy part of the theorem. Suppose X has an exponential distribution with parameter λ. Then we have

$$\mathbb{P}\left[X \geq t\right] = 1 - F_X(t) = e^{-\lambda t}.$$

Hence,

$$\mathbb{P}\left[X \geq s + t | X \geq s\right] = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}\left[X \geq t\right].$$

On the other hand, let us assume that X is a non-negative continuous random variable which is memory-less. Let $h(t) = \mathbb{P}\left[X \geq t\right]$. The assumption that X is memory-less can be written in the form:

$$h(s + t) = h(s)h(t).$$

Assuming that $h(1) = a$, it is easy to see that $h(n) = a^n$ for every $n \geq 1$, and moreover, for any rational number r, we have $h(r) = a^r$. The continuity of h, then shows that $h(x) = a^x = e^{-\lambda x}$, where $\lambda = -\log a$. This shows that $F_X(t) = 1 - e^{-\lambda t}$ and hence $f_X(t) = \lambda e^{-\lambda t}$ for all $t \geq 0$.

□

Example 3.4.5 (Gaussian distribution)**.** Normal or Gaussian random variables are some of the most important examples of continuous random variables. They arise naturally in the central limit theorem. In many practical situations, when a quantity is the sum of a large number of independent identically distributed quantities, Gaussian variables emerge as a result of this.

Definition 3.4.9. A continuous random variable X is said to have Gaussian or normal distribution with parameters (μ, σ^2), written $X \sim N(\mu, \sigma^2)$, if the probability density function of X is given by

$$f_X(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

A random variable with normal distribution with parameters $\mu = 0$ and $\sigma = 1$ is called a *standard normal distribution*.

Remark 3.4.10. Note that as opposed to the previous examples of continuous random variables, a Gaussian random variables takes values in the entire set of real numbers. In spite of the importance of the Gaussian variables, there is no closed formula for the distribution function of such variables. The distribution function of the standard normal distribution is often denoted by $\Phi(t)$.

The following theorem will be useful later:

Proposition 3.4.11 (Standardization). *Let $X \sim N(\mu, \sigma^2)$ and $Y = \frac{X-\mu}{\sigma}$, then Y has standard normal distribution, that is $Y \sim N(0, 1)$.*

Proof. This is a straightforward computation. □

Example 3.4.6. [log-normal Distribution] The log-normal distribution plays an important role in the so-called Black-Scholes formula in financial mathematics. We will first give the definition:

Definition 3.4.12. A random variable X is said to have *log-normal* distribution if $\log X$ has a normal distribution. In other words, X has the log-normal distribution with parameters μ and σ if $X = e^Y$ where $Y \sim N(\mu, \sigma^2)$.

Proposition 3.4.13. *The probability density function of a log-normal distribution with parameters μ and σ is given by*

$$f_X(t) = \begin{cases} \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

Proof. Assume that Z is a standard normal distribution, $Y = \mu + \sigma Z$ and $X = e^Y$. We have

$$F_X(t) = \mathbb{P}[X \leq t] = \mathbb{P}[\exp(Y) \leq t] = \mathbb{P}[Y \leq \log t] = \mathbb{P}\left[Z \leq \frac{\log t - \mu}{\sigma}\right] = \Phi\left(\frac{\log t - \mu}{\sigma}\right).$$

Differentiating this, we have

$$f_X(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right).$$

□

In the same way that a normal distribution is the average of many independent centered Bernoulli distributions, a lognormal can be viewed as the a geometric average of a large number of multiplicatively centered ($\mathbb{E}[\log X] = 0$) Bernoulli distributions.

4
Expectation

4.1 What is Expectation?

Let X be a discrete or continuous random variable. Various numerical quantities may provide partial information about the distribution of X. One of most important ones is the expected value of X. We will give the definition for discrete and continuous random variables separately.

Definition 4.1.1. For a discrete random variable X with values $x_1, x_2, \ldots$, the *expected value* of X is defined by

$$\mathbb{E}[X] = \sum_i x_i \cdot \mathbb{P}[X = x_i].$$

Remark 4.1.2. Suppose X only takes a finite number of values. The expected value as defined above is then a finite sum, hence unambiguously defined. However, when X takes infinitely many values, the sum may be ambiguous. To see why, recall that if $a_1, a_2, \ldots$ is a sequence of real numbers, then the value of the sum

$$\sum_{n=1}^{\infty} a_n$$

may depend on the ordering of the terms. As the values of a random variables are not given in any particular order, it may happen that different rearragements of the terms in

$$\sum_{i=1}^{\infty} x_i \cdot \mathbb{P}[X = x_i].$$

lead to different values. This can be avoided by assuming that the series is *absolutely convergent*. In other words, the expected value of X is defined if and only if the series

$$\mathbb{P}[X] = \sum_i |x_i| \cdot \mathbb{P}[X = x_i].$$

is convergent.

Remark 4.1.3. The above sum can be viewed as a weighted sum of the values of the random variable X. The weight given to each possible value x_i is the probability that is its attained by X.

Example 4.1.1. Let X be a Bernoulli random variable with parameter p. Then

$$\mathbb{E}[X] = p \cdot 1 + (1-p) \cdot 0 = p.$$

Example 4.1.2. Let X be a Binomial random variable with the parameters (n, p). Then

$$\mathbb{E}[X] = \sum_{j=0}^{n} j \binom{n}{j} p^j (1-p)^{n-j} = np.$$

Example 4.1.3. Let X be a discrete random variable with Poisson distribution with parameter λ. Then,

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

Expected value of positive integer-valued random variables

Let X be a random variable that only takes only non-negative integer values. Examples include any random variable that counts the *number* or *cardinality* which is random. For such random variables, one can give an alternative formula for the expectation that proves more handy in some cases.

Theorem 4.1.4. *Suppose X is a random variable which only takes values $0, 1, 2, \ldots$. Then the expected value of X is given by*

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}[X \geq i].$$

Proof. We will unpack the right-hand side of the equation

$$\begin{aligned} \sum_{i=1}^{\infty} \mathbb{P}[X \geq i] &= \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} \mathbb{P}[X = j] = \sum_{j=1}^{\infty} j \cdot \mathbb{P}[X = j] \\ &= \sum_{j=0}^{\infty} j \cdot \mathbb{P}[X = j] = \mathbb{E}[X]. \end{aligned} \tag{4.1}$$

□

Example 4.1.4. Let X have the geometric distribution with parameter p, i.e.

$$\mathbb{P}[X = j] = p(1-p)^{j-1},$$

for $j \geq 1$. Compute $\mathbb{E}[X]$.

It is instructive to first try a direct approach. Using the definition of expectation, we can write

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} ip(1-p)^{i-1}.$$

However, it is not evident how this series can be summed. A better approach is to use Theorem 4.1.4. Since $\mathbb{P}(X \geq i) = (1-p)^{i-1}$, we will have

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} (1-p)^i = \frac{1}{p}.$$

Example 4.1.5. A fair die is rolled. Suppose that X is the number shown. Compute $\mathbb{E}[X]$.

Clearly X takes value $1 \leq n \leq 6$, each with probability $1/6$. From here we have:

$$\mathbb{E}[X] = \frac{1}{6}\sum_{n=1}^{6} n = \frac{21}{6} = 3.5$$

Example 4.1.6. Consider the Poisson random variable with parameter λ. Compute $\mathbb{E}[X]$.

Note that X takes positive integers as its values; we have $\mathbb{P}(X = i) = e^{-\lambda}\frac{\lambda^i}{i!}$. So, we can apply the definition of the expected value to get:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} \cdot i = \sum_{i=1}^{\infty} e^{-\lambda}\frac{\lambda^i}{(i-1)!} = \lambda$$

4.2 Expected Value of Continuous random variables

Let X be a continuous random variable with density function $f_X(t)$ and distribution function $F_X(t)$. Our goal is to define the expected value of X, as we defined it for discrete random variables. Again, we are guided by the physical definition of the center of mass. In the discrete case, the mass is distributed as points, here the mass is continuously distributed across the real line. To find the center of mass, one has to average up all the points with respect to the weights given to them. From here, the following definition seems very natural:

Definition 4.2.1. Let X be a continuous random variable with the density function $f_X(t)$. The expected value of the mean of X is defined by

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(t) dt.$$

Remark 4.2.2. Note that for the above integral to be defined, we require the absolute convergence of the integral. In other words, one must have

$$\int_{-\infty}^{\infty} |x| f_X(t) dt < \infty.$$

Example 4.2.1. Let X have uniform distribution over the interval $[a, b]$. Then, we have

$$\mathbb{E}[X] = \int_a^b \frac{x}{b-a} dt = \frac{a+b}{2}.$$

Note that the answer is consistent with the idea that the expectation corresponds to the center of mass.

Example 4.2.2. Let X have the Cauchy distribution given by

$$f_X(t) = \frac{1}{\pi(1+t^2)}.$$

It is easy to see that the integral

$$\int_0^{\infty} \frac{t}{\pi(1+t^2)} dt$$

is divergent, hence $\mathbb{E}[X]$ is not defined.

We have an analogue of Theorem 4.1.4 for continuous random variables.

Theorem 4.2.3. *Let X be a non-negative continuous random variable with the distribution function $F_X(t)$. Then*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X \geq t]\, dt = \int_0^\infty (1 - F_X(t))dt.$$

Let X be a random variable, $f : \mathbb{R} \to \mathbb{R}$ a continuous function and $Y = h(X)$. Note that Y is a random variable whose value depends on X. In various situation, we are interested in computing $\mathbb{E}[Y]$. The direct way involves computing the density function of Y and using the definition of expected value. The following theorem provides a shortcut:

Theorem 4.2.4. *Let X be a continuous random variable with the density function $f_X(t)$. For any continuous function $h(t)$, we have*

$$\mathbb{E}[h(X)] = \int_{-\infty}^\infty h(t) f_X(t) dt.$$

Example 4.2.3. Let X have the uniform distribution over the interval $[0, \pi]$. Compute $\mathbb{E}[\sin X]$.

We have

$$\mathbb{E}[\sin X] = \int_0^\pi \frac{1}{\pi} \sin t dt = \frac{2}{\pi}.$$

Before doing any computations, we first list some of the most basis properties of expectation: