An introduction to probability and statistics

Edoardo Provenzi

Contents

Introduction

What is probability and why is it useful? These are the two most important questions that we will try to answer in this course.

Roughly speaking, probability is the mathematical systematization of the study of chance in random phenomena. There are different kinds of such phenomena:

- the roll of a die and the selection of a card from a well shuffled deck of 52 cards are common examples of unconditioned random phenomena, because they are not subjected to any constraint
- the height of a person is of course random, but it may depend on the environment or on genetic factors, which, if taken into account, may provide constraints that change the computation of the probability that a person has a certain height.

Without probability it is practically impossible to grasp the correct interpretation from data analysis through modern statistical methods. For this reason, it is quite natural to study probability prior to studying statistics.

Probability allow us to quantify the strength or 'confidence' on an hypothesis that we have guessed about the behavior of a given phenomenon.

Finally, the methods and results of probability provides the transition from descriptive statistics to inferential methods, i.e. techniques that allow us not only to describe a phenomenon but to predict its future behavior.

The course is developed following a bottom-up pedagogical choice, which starts with the discussion of motivational examples and draws general conclusion afterwards, a technique that can be summarized as '*the concrete comes before the abstract*'. In particular, I tried to *avoid as much as possible unmotivated definitions* with the hope that this choice will help the reader during the learning process.

The author.

Chapter 1

A gentle introduction to probability

We start with a little bit of history. The legend says that the theory of probability was born in the seventeenth century when Chevalier de Méré, a writer with a passion for gambling, was losing quite a lot of money due to his wrong assumptions about the chances of getting a double 6 in a roll of two dice. He asked his friends Blaise **Pascal** (1623-1662) and Pierre de **Fermat** (1607-1665) to help him by developing a mathematical formulation of gambling odds that would improve his chances at winning. Incredibly enough, this led to the development of the study of probability!

Probability was formalized as a rigorous mathematical theory by the great Russian mathematician Andrey **Kolmogorov** (1903-1987), 300 years after its casual birth. One of the reasons why it took so long to achieve this accomplishment is that this formalization needs both set theory, developed by Georg **Cantor** (1845-1918) at the end of the nineteenth century, and the abstract theory of measure and integration developed by Herni **Lebesgue** (1875-1941), at the beginning of the twentieth century.

The powerful methods of statistics that are used nowadays in practically every discipline are grounded on probability, hence we have this chain of inferences:

$$\text{(Set theory + Measure and Integration)} \implies \text{Probability} \implies \text{Statistics.}$$

In this course, we will minimize the mathematical formalism in favor of concrete examples and practical applications. However, in order to acquire at least the basic terminological foundation and mathematical concepts necessary to tackle engaging examples and problems, a brief overview of the essential mathematical tools of probability is indispensable.

To avoid the common pitfall of overwhelming learners with numerous definitions and results presented all at once, detached from practical context, we will introduce the mathematics of probability incrementally, integrating worked examples at each step.

1.1 The peculiar nomenclature of probability

A first obstacle in studying probability is the particular vocabulary used in this discipline, which, nevertheless, must be learned.

The first item to be defined is the type of experiments of interest in probability.

Def. 1.1.1 (Random experiment) *An experiment that, reproduced in exactly the same way, may lead to different outcomes which cannot be known in advance is called random.*

On the contrary, in a **deterministic experiment**, once we set up the same initial conditions, we measure the same outcome for every repetition.

Def. 1.1.2 (Sample (or probability) space) *The set*[1] *S of all possible outcomes of a random experiment is called the sample (or probability) space for the experiment.*

Example 1.1.1 If the experiment consists in tossing a coin, then

$$S = \{H, T\},$$

where H stays for 'head' and T for 'tail'. ◇

Example 1.1.2 If the experiment instead is the reaction time t of a puppy after being called, then

$$S = (0, +\infty),$$

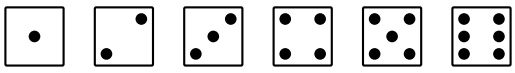
because the puppy can react almost instantly, so $t \to 0$, or not at all, in which case $t \to +\infty$. ◇

These two examples show an important differences between sample spaces: S **can be either discrete or continuous, finite or infinite**.

We will see that the distinction between sample spaces turns out to be important because it determines the way in which probabilities can be assigned.

Once the sample space has been defined, we can consider collections of possible outcomes of an experiment, as shown in the next example.

Example 1.1.3 Imagine to roll a die, then the sample space is



or, in mathematical symbols, $S = \{1, 2, 3, 4, 5, 6\}$.

One of the most natural question that we can ask when we roll a die is whether the outcome will be *even* or *odd*, independently of the numerical value. In either one or the other case, the outcome will belong to one of the following two subsets of S:

$$E := \{2, 4, 6\} \subset S,$$

$$O := \{1, 3, 5\} \subset S.$$

◇

[1] Another letter very often used instead of S is Ω.

This example motivates why, in probability, it is convenient to introduce a name for *a collection* of possible outcomes and *not only* a single outcome: this gives us the possibility to 'group together' outcomes that share a common property. The name that has been chosen is 'event'.

Def. 1.1.3 (Event) *An event is any collection of possible outcomes, i.e., any subset of S, including S itself and the empty set.*

Def. 1.1.4 *An event $E \subset S$* ***occurs*** *if the outcome of a random experiment belongs to E.*

Example 1.1.4 Using the notation of the previous example: if we roll the die and the outcome is 3, then the event O occurs because $3 \in O$. Instead, if we obtain 4, then the event E occurs because $4 \in E$. ◇

Since events are subsets of the sample space, it should not come as a surprise that the so-called **event operations** are simply the typical transformations that we can perform over the subsets of a given set.

- **Union** $\cup$: $E \cup F$ is the set that contains the elements of E and those of F, without repetitions. For instance, if $E = \{1, 2, 3, 4\}$ and $F = \{3, 5, 6\}$, then $E \cup F = \{1, 2, 3, 4, 5, 6\}$.
- **Intersection** $\cap$: $E \cap F$ is the set that contains only the elements shared by E and F. Considering again the previous sets, we have $E \cap F = \{3\}$.
- **Complementary**: this last operation consists in considering the elements of S which *are not* in the event E, the resulting event is denoted by E^c. For example, if we consider the sample space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and E, F are as above, then $E^c = \{5, 6, 7, 8, 9, 10\}$ and $F^c = \{1, 2, 4, 7, 8, 9, 10\}$.

The **probabilistic interpretation of the event operation** is the following.

- **Union**: *at least one* event occurs.
- **Intersection**: *both* events occur *simultaneously*.
- **Complementary**: the *contrary* of the event occurs.

If the intersection of two event is empty, then, given the probabilistic interpretation of intersection, the two events cannot occur simultaneously.

Def. 1.1.5 (Disjoint (or mutually exclusive, or incompatible) events) *Two events E and F are disjoint (or mutually exclusive, or incompatible) if $E \cap F = \varnothing$. More generally, multiple events $E_1, E_2, \ldots, E_n$ have this property if $E_i \cap E_j = \varnothing$ for all $i \neq j$.*

Example 1.1.5 The events E and O of Example 1.1.3 are disjoint:

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

◇

Two very famous laws relate the operations of union, intersection and complementary, they are called **De Morgan laws**: given events $E, F \subseteq S$ we have

$$(E \cup F)^c = E^c \cap F^c, \quad (E \cap F)^c = E^c \cup F^c,$$

so, the *complementary reverses the role of union and intersection.*

In the sequel, we will also need to use the **distributive relations** between union and intersection:

$$E \cap (F \cup G) = (E \cap F) \cup (E \cap G), \text{ more generally, } E \cap \bigcup_{j=1}^{n} E_j = \bigcup_{j=1}^{n} E \cap E_j, \tag{1.1}$$

$$E \cup (F \cap G) = (E \cup F) \cap (E \cup G), \text{ more generally, } E \cup \bigcap_{j=1}^{n} E_j = \bigcap_{j=1}^{n} E \cup E_j. \tag{1.2}$$

The last definition of the typical vocabulary of probability that we must introduce is that of partition. To motivate this concept let us consider again the act of rolling a die: the events E and O are not only disjoint, but their union reconstructs the sample space, in fact

$$E \cup O = \{2, 4, 6\} \cup \{1, 3, 5\} = \{1, 2, 3, 4, 5, 6\} = S.$$

E and O are particular instances of what is called a partition of a sample space.

Def. 1.1.6 (Partition of the sample space) *A collection of events $E_j \subset S$ that*

- *are disjoint: $E_i \cap E_j = \varnothing$ for all $i \neq j$*
- *reconstruct the sample space via union, i.e. $\bigcup_j E_j = S$*

is said to form a partition of S.

To underline the fact that in a partition the union is performed over disjoint subsets of S, the symbol $\bigcup$ is often replaced by $\bigsqcup$. We will adopt this useful convention in the rest of the course, so:

$$E \sqcup F := \text{union of two disjoint sets } E \text{ and } F,$$

and more generally,

$$\bigsqcup_j E_j := \text{union of disjoint sets } E_j.$$

1.2 The probability function

Imagine that an experiment is performed n times in an identical way and that some outcomes happen more often than others. This permits to define a non-trivial '*frequency of occurrence*' of events as follows: if we indicate with $\sharp(E)$ the amount of times that the event E occurs, then its frequency of occurrence is

$$f(E) = \frac{\sharp(E)}{n}.$$

For centuries, the probability associated to an event was defined through the following limit:

$$\mathscr{P}(E) = \lim_{n \to +\infty} \frac{\sharp(E)}{n}.$$

Despite of being very intuitive, this approach has some serious drawbacks that are far from being only of theoretical nature. For instance: does the limit exist? If so, how can we practically compute it? Namely, how many experiments do we have to run in order to have a fair estimation of the frequency of E?

In 1933, Kolmogorov published the book *Foundations of the Theory of Probability*, which allowed us to overcome the problems related to the frequency approach at the expense of a more abstract interpretation of probability.

The basic concept underlying Kolmogorov approach is the *probability function* P, which is a map that takes *events as input* and gives back *values in the set* $[0, 1]$ *as output*.

The numbers between 0 and 1 must be interpreted as probabilities, for examples:

- 0 is an impossible event
- $1/2$ is an event that has 50% of probability to occur
- 1 is an event which occur with a 100% probability, i.e. a certain event

Before giving the definition of P, let us analyze an example that will allow us understanding what kind of properties such a function should have.

Example 1.2.1 Consider again the experiment of tossing a *fair coin*, i.e. a balanced coin that is equally as likely to land heads up as tails up. We already know that the sample space is $S = \{H, T\}$ and our intuition suggests that a reasonable probability function should assign equal probabilities to heads and tails, i.e. it must satisfy $P(H) = P(T)$. Moreover:

1. $S = H \cup T$, so we have $P(H \cup T) = 1$, because we have 100% certainty to obtain either head or tail
2. $H^c = T$ and $T^c = H$, hence the set on which the probability function is defined should also contain the complementary sets of the events
3. $H \cap T = H \cap H^c = \varnothing = T \cap H = T \cap T^c$, and since it is not possible to obtain both head and tail simultaneously, we must have $P(\varnothing) = 0$
4. from item 3, it follows that H and T are disjoint and from item 1 it follows that H and T reconstruct S through their union, so $\{H, T\}$ is a partition of S. The fact that $P(H) = P(T)$ and $P(H \sqcup T) = 1$ implies that $P(H) = P(T) = 1/2$ and $P(H \sqcup T) = P(H) + P(T)$. $\diamond$

The following definition of probability function is the natural generalization of this example.

Def. 1.2.1 (Probability function) *Given a sample space S, a probability function (or measure) is a function*

$$\begin{array}{rccc} P: & \mathcal{A} & \longrightarrow & [0,1] \\ & E & \longmapsto & P(E), \end{array}$$

where $\mathcal{A}$ is a set containing:

- *all the events $E \subseteq S$ and their complementary sets E^c*
- *their unions $\bigcup_j U_j$ and their intersections $\bigcap_j U_j$,*

*and P satisfies the so-called **Kolmogorov axioms**:*

K1. $P(S) = 1$.

*K2. For any set of **disjoint events** $E_j \subset S$ it holds that*

$$P\left(\bigsqcup_j E_j\right) = \sum_j P(E_j).$$

Let us now show that, despite its abstraction, when the Kolmogorov approach descends on Earth and we applied to real-world problems, it agrees with our intuition.

Example 1.2.2 Consider again a die with its sample space $S = \{1, 2, 3, 4, 5, 6\}$. If no face of the die is different than the others, we can assume that the probability of the event $\{n\}$, with $n = 1, \ldots, 6$, is a constant value $p \in [0, 1]$ which does not depend on the number on the face. Axiom K1 implies that $P(S) = 1$, but

- $S = \{1\} \cup \{2\} \cup \{3\} \cup \{4\} \cup \{5\} \cup \{6\}$
- $\{m\} \cap \{n\} = \varnothing$ for all $m, n = 1, \ldots, 6$, $m \neq n$,

so $S = \{1\} \sqcup \{2\} \sqcup \{3\} \sqcup \{4\} \sqcup \{5\} \sqcup \{6\}$ is a partition of S and axiom K2 implies

$$P(S) = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) + P(\{6\}) = p + p + p + p + p + p = 6p,$$

so $P(S) = 1$ if and only if $6p = 1$, i.e. $p = 1/6$.

Of course, the usefulness of the abstraction introduced by Kolmogorov reveals itself when we deal with more complicated problems, however it is reassuring to 'touch with our hands' that it works perfectly also for problems as simple as the rolling of a die. ◇

It would be of course very useful to have a general methods of defining legitimate probability functions without having to check Kolmogorov's axioms.

A response to this need is given by the following result.

Theorem 1.2.1 *Let $S = \{s_1, \ldots, s_n\}$ be a finite sample space. If $(p_j)_{j=1}^n \in [0,1]$ is a finite sequence of numbers such that*

$$\sum_{j=1}^{n} p_j = 1, \tag{1.3}$$

then the following

$$\begin{array}{rccl} P: & \mathcal{A} & \longrightarrow & [0,1] \\ & E & \longmapsto & P(E) := \displaystyle\sum_{\{k\,:\,s_k \in E\}} p_k, \end{array}$$

is a probability function.

Notice that:

$$\sum_{\{k\,:\,s_k \in E\}} \text{ means that the sum is performed over the indices } k \text{ such that } s_k \in E.$$

Instead of going through the proof, let us discuss a practical example that will help us understanding how to interpret and use this theorem.

Example 1.2.3 Consider a game of darts where, as usual, a player throwing a dart at a board receives a score corresponding to the number assigned to the region in which the dart lands. Referring to Figure 1.1, we consider a dart board with radius r and distance between rings given by $r/5$.

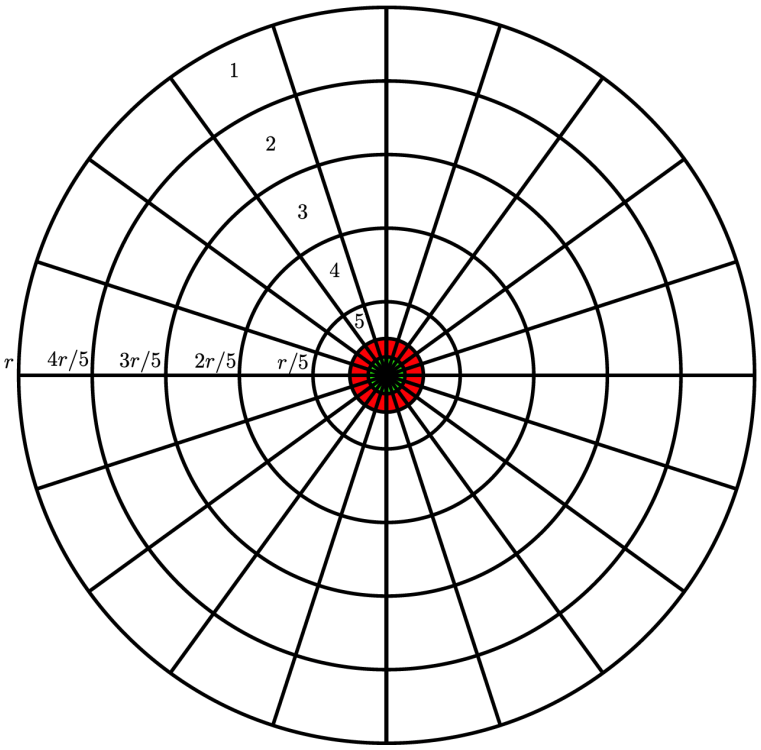


Figure 1.1: The dart board considered in this example.

The sample space here is the finite set $S = \{1, 2, 3, 4, 5\}$. For a *non-expert player*, it *seems reasonable to assume that the probability* of the dart hitting a particular region *is proportional to the area* of that region. Thus, a bigger region has a higher probability of being hit.
If we make the assumption that the board is always hit, then we have

$$p_k := P(\text{scoring } k \text{ points}) = \frac{\text{Area of the region } k}{\text{Area of the board}}.$$

The areas of the regions corresponding to 1 point and 5 points are the only ones that can be computed directly, so let us start with them.

- Region 1: its area is the difference between the total area of the board, i.e. πr^2, and the area of the board until the radius $4r/5$:

$$\text{Area of region } 1 : \pi r^2 - \pi r^2 \left(\frac{4}{5}\right)^2 = \pi r^2 \left(1 - \left(\frac{4}{5}\right)^2\right).$$

- Region 5:

$$\text{Area of region } 5 : \pi r^2 \left(\frac{1}{5}\right)^2.$$

It follows that

$$p_1 = \frac{\pi r^2 \left(1 - \left(\frac{4}{5}\right)^2\right)}{\pi r^2} = 1 - \left(\frac{4}{5}\right)^2 = 0.36, \qquad p_5 = \frac{\pi r^2 \left(\frac{1}{5}\right)^2}{\pi r^2} = \left(\frac{1}{5}\right)^2 = 0.04.$$

Hence, P assigns 36% of probability of scoring 1 point and 4% of scoring 5 points to a beginner.

The areas of the k-th regions, $k = 2, 3, 4$, are given by the difference between the total area, i.e. πr^2, and the sum of the area of the external and internal regions with respect to the k-th one.

- k-th internal region:

$$\text{int}(k) = \pi r^2 \frac{(5-k)^2}{5^2}, \quad k = 2, 3, 4.$$

- k-th external region: total area of the board minus area until k, which is $\text{int}(k)$ plus the area of the k-th region:

$$\text{ext}(k) = \pi r^2 - \pi r^2 \frac{(5-k+1)^2}{5^2} = \pi r^2 - \pi r^2 \frac{(6-k)^2}{5^2}, \quad k = 2, 3, 4.$$

So:

$$\text{Area region } k = \cancel{\pi r^2} - \left[\pi r^2 \frac{(5-k)^2}{5^2} + \cancel{\pi r^2} - \pi r^2 \frac{(6-k)^2}{5^2}\right] = \pi r^2 \frac{(6-k)^2 - (5-k)^2}{5^2},$$

which implies:

$$p_k = \frac{\pi r^2 \frac{(6-k)^2 - (5-k)^2}{5^2}}{\pi r^2} = \frac{(6-k)^2 - (5-k)^2}{5^2}, \quad k = 2, 3, 4.$$

We notice the previous formula gives the correct probability also for $k = 1$ and $k = 5$.

By theorem 1.2.1, P is a probability function if the values p_k sum to 1:

$$\sum_{k=1}^{5} p_k = \frac{(5^2 - 4^2) + (4^2 - 3^2) + (3^2 - 2^2) + (2^2 - 1) + 1}{5^2} = \frac{5^2}{5^2} = 1.$$

This shows that our initial guess is correct, i.e. the fraction between the area hit by the dart and the total area of the board gives rise to a probability function.

Consequently, the probability to hit the different regions under the initial hypothesis are

$$\begin{cases} p_1 = 36\% \\ p_2 = 28\% \\ p_3 = 20\% \\ p_4 = 12\% \\ p_5 = 4\%. \end{cases}$$

So, for a beginner, the probability of hitting decreases of 8% as we move from one region to the next internal one.

Figure 1.2 shows these probabilities as histograms.

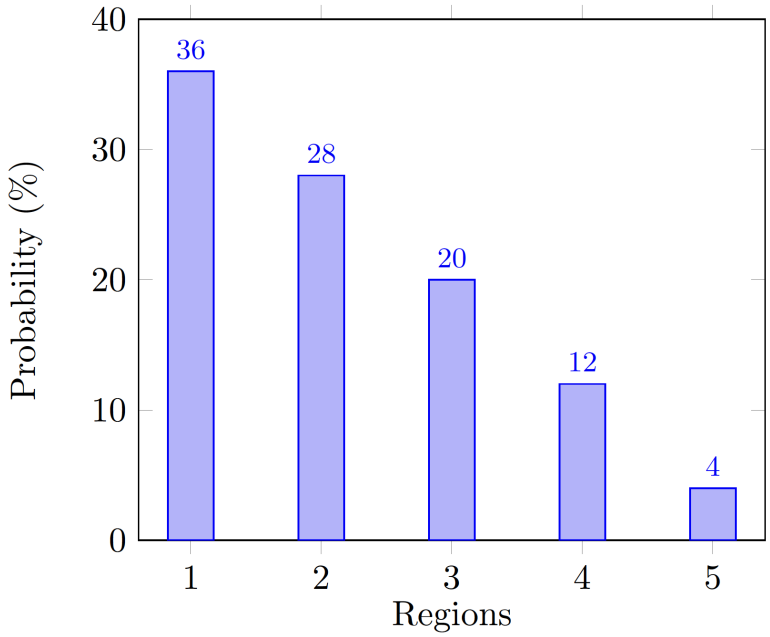


Figure 1.2: Histogram of probabilities.

◇

1.2.1 Properties of the probability function

To agree with common sense, Kolmogorov's definition of probability function should assign the probability $1 - P(E)$ to E^c and 0 to the empty set $\varnothing$. The following proposition states that this is indeed the case.

Theorem 1.2.2 *For all event $E \in \mathcal{A}$, any probability function P has the following property $P(E^c) = 1 - P(E)$. Thus, in particular, $P(\varnothing) = 0$.*

Proof. Given an event $E \in \mathcal{A}$ we have:

- $E \cap E^c = \varnothing$
- $E \sqcup E^c = S$,

so we can write a partition of the sample space as follows $S = E \sqcup E^c$, then

$$1 \underset{K1}{=} P(S) = P(E \sqcup E^c) \underset{K2}{=} P(E) + P(E^c),$$

it follows that $P(E^c) = 1 - P(E)$. In particular, if $E = S$ we have $P(S^c) = P(\varnothing) = 1 - P(S) = 0$. □

Example 1.2.4 Consider the sample space of a deck of cards, i.e. $S = \{C, D, H, S\}$, and the event $E = \{C, H\}$, then $E^c = \{D, S\}$. The event E corresponds to picking a card from the deck that is either clubs or hearts, while the event E^c is picking a card that is either diamonds or spades. Since half of the cards belong to E and the other half belongs to E^c, it makes sense to set $P(E) = P(E^c)$, but then the previous theorem implies

$$P(E^c) = P(E) = 1 - P(E) \iff 2P(E) = 1 \iff \frac{1}{2} = P(E) = P(E^c),$$

i.e. we have equal 50% probability of picking a card that is either clubs or heart, on one side, or of picking a card that is either diamonds or spades, on the other side. This, of course, agrees with common sense. ◇

Another *seemingly* intuitive properties of the probability function is that, if an event E is composed by less outcomes than another one F, i.e. $E \subseteq F$, then the probability assigned to E should be smaller than that assigned to F. This is indeed the case, but to prove it we must first prove a technical properties of the probability function P.

Theorem 1.2.3 *If $P : \mathcal{A} \to [0, 1]$ is a probability function and E and F are any events in $\mathcal{A}$, then:*

1. $P(F \cap E^c) = P(F) - P(E \cap F)$
2. *If $E \subseteq F$, then $P(E) \leqslant P(F)$.*

Proof.

1. To prove the first statement we simply have to consider the Venn diagram depicted in Figure 1.3.

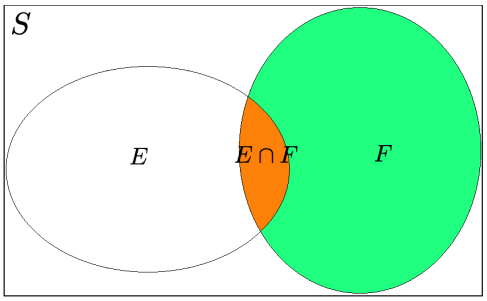


Figure 1.3: The Venn diagram useful to prove property 1.

From this it follows that the identity $F = (E \cap F) \sqcup (F \cap E^c)$ holds true.

So, from axiom $K2$ we have $P(F) = P(E \cap F) + P(F \cap E^c)$, or, rearranging the terms, $P(F \cap E^c) = P(F) - P(E \cap F)$.

2. Now, looking at the the Venn diagrams in Figure 1.4,

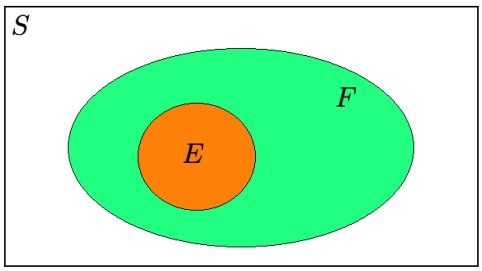


Figure 1.4: The Venn diagram useful to prove property 3.

we see that if $E \subseteq F$, then $E \cap F = E$, so, using property 1. and the fact that $P(E)$ is always $\geqslant 0$, we have

$$0 \leqslant P(F \cap E^c) \underset{1.}{=} P(F) - P(E \cap F) = P(F) - P(E) \iff P(E) \leqslant P(F).$$

□

We introduced the implication $E \subseteq F \implies P(E) \leqslant P(F)$ by declaring that it is seemingly intuitive, in the next example we explain why we added 'seemingly'...

Example 1.2.5 (**The Linda problem**) This problem was introduced by the cognitive psychologists Amos **Tversky** and Daniel **Kahneman** in 1981 and goes like this.

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which event more is probable?

1. *Linda is a bank teller.*

2. *Linda is a bank teller and is active in the feminist movement.*

The great majority of people that answered this question chose the answer 2. declaring that the description of Linda was more likely to be that of a feminist than that of a bank teller.

However, if we consider the set of all female bank tellers and that of all female *and* feminist bank tellers, the second one is clearly contained in the first: every female *and* feminist bank teller is a bank teller, while it is not necessarily true that a female bank teller is also feminist.

This explains why, in probability, even a seemingly intuitive implication may be wrongly interpreted. ◇

By the axiom $K2$, the probability of the disjoint union of two events is the sum or their corresponding probabilities, but what happens if the union is not disjoint, i.e. if the two events intersects non-trivially?

In this case, the sum of the probability of each event takes into account two times the outcomes belonging to the intersection.

So, in order to treat the outcomes belonging to the intersection as all the others outcomes, the probability function of the union of two events should be the sum of the probability of the events minus the probability of theirs intersection.

This is exactly what happens.

Corollary 1.2.1 *Given a probability function $P : \mathcal{A} \to [0, 1]$ and two events $E, F \in \mathcal{A}$, we have:*

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) \tag{1.4}$$

Proof. Looking at the the Venn diagrams in Figure 1.5,

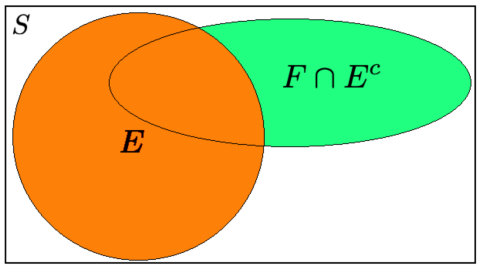


Figure 1.5: The Venn diagram useful to prove this proposition.

it follows that $E \cup F = E \sqcup (F \cap E^c)$, so

$$P(E \cup F) = P(E) + P(F \cap E^c) \underset{\text{1. of th. 1.2.3}}{=} P(E) + P(F) - P(E \cap F).$$

□

The formula of the previous proposition implies that $P(E \cup F)$ is smaller than the sum of the probabilities of E and F, because we subtract to this sum the positive value $P(E \cap F)$, so

$$P(E \cup F) \leqslant P(E) + P(F).$$

This fact can be generalized thanks to the famous **Booles' inequality**:

$$P\left(\bigcup_{j=1}^{n} E_j\right) \leqslant \sum_{j=1}^{n} P(E_j),$$

which holds for any collection $E \in \mathcal{A}$, $(E_j)_{j=1}^n$ of events in $\mathcal{A}$.

In the example 1.3.3 we will show the usefulness of the previous theorem in a practical case.

We close this section with a result that will be very useful later on (in the discussion of the Bayes' theorem) which shows the benefits of the Kolmogorov axiomatization of probability through set theory: in fact, the proof of this result is relatively easy, while it would be difficult to obtain with other methods.

Theorem 1.2.4 *Let $P : \mathcal{A} \to [0, 1]$ be a probability function,*

- $E \in \mathcal{A}$
- $(F_k)_{k=1}^n$ *a partition of the sample space S.*

Then:

$$P(E) = \sum_{k=1}^{n} P(E \cap F_k). \tag{1.5}$$

Before providing the proof, let us interpret the statement by considering only two terms. If $\{F_1, F_2\}$ is a two-set partition of S, i.e., $F_1 \cap F_2 = \varnothing$ and $F_1 \cup F_2 = S$, then eq. (1.5) implies

$$P(E) = P(E \cap F_1) + P(E \cap F_2),$$

this means that we can recover the probability of the event E by adding the partial probabilities of the simultaneous events $E \cap F_1$ and $E \cap F_2$. We will see in the proof that the validity of this fact relies heavily on the properties of a partition.

Proof. Since $E = E \cap S$ and $S = \bigsqcup_{k=1}^{n} F_k$, then, by the distributivity between intersection and union we have

$$E = E \cap \bigsqcup_{k=1}^{n} F_k \underset{(1.1)}{=} \bigsqcup_{k=1}^{n} E \cap F_k,$$

applying P to the first and the last member of the previous equality we get

$$P(E) = P\left(\bigsqcup_{k=1}^{n} E \cap F_k\right),$$

but the events F_k are all disjoints, so also the events $E \cap F_k \subseteq F_k$ are, and thus, by axiom $K2$ we find

$$P(E) = \sum_{k=1}^{n} P(E \cap F_k).$$

□

Before continuing our analysis of probability, we need to introduce the basic fact about combinatorial calculus.

1.3 Enumerating outcomes: a basis of combinatorial calculus

Combinatorics is 'the art' of counting, a process of that can become quite sophisticated when placed in the hands of a probabilist or a statistician, who must treat problems subject to many restrictions. The way to solve such problems is to break them down into a series of simple tasks that are easy to count, and employ known rules of combining tasks. The following theorem is a first step in such a process.

Theorem 1.3.1 (Fundamental Theorem of Counting) *If a job consist of k separate tasks, the i-th of which can be done in n_i ways, $i = 1, \ldots, k$, then the entire job can be done in $n_1 n_2 \cdots n_k$ ways.*

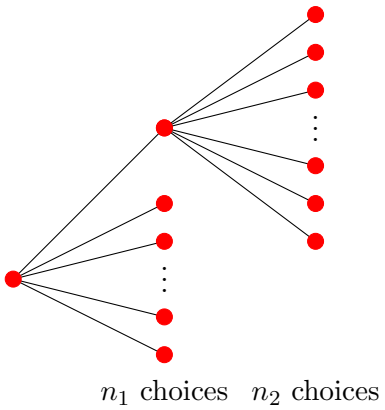
Proof. The general idea of the proof can be understood already for $k = 2$ and then generalized by induction. So, we will only deal with the case $k = 2$.

The first task can be done in n_1 ways and, for each of these ways, we have n_2 choices for the second task. Thus, the job can be done in this number of ways:

$$\underbrace{(1 \cdot n_2) + (1 \cdot n_2) + \cdots + (1 \cdot n_2)}_{n_1 \text{ terms}} = n_1 \cdot n_2,$$

thus proving the theorem for $k = 2$. □

A very intuitive visualization of the theorem is given by the following diagram (courtesy of Bernhard Haak).



Example 1.3.1 A typical example that shows how the fundamental theorem of counting can be used in practice, and also the care that one must have when using it, is given by *the lottery*.

Imagine that we can extract 50 different numbers, then we have to distinguish between two types of lotteries.

- The most common is the *lottery without replacement*, i.e. once a number is extracted, it is removed from the set of numbers that may appear in the following extraction. So, one number is allowed to be extracted only once.
- However, there exists also a *lottery with replacement*, in which each number that has been extracted is set back to the ensemble of number that may appear in the following extraction. This means that the same number is allowed to be extracted multiple times.

In order to be coherent with the proof, that we have performed only in the case $k = 2$, let us define our job to be: 'extract 2 numbers'. Then,

- for a *lottery without replacement*, the job can be done in $50 \cdot 49 = 2450$ ways, because there are 50 possible outcomes in the first extraction and 49 in the second one;
- for a *lottery with replacement*, the job can be done in $50 \cdot 50 = 2500$ ways, because, having reintroduced the number extracted the first time in the original ensemble, there are 50 possible outcomes in the both first and the second extraction. ◇

Besides replacement, there is another aspect that must be taken into account when considering counting, which is *ordering*.

We can illustrate this aspect using again the example of lottery: imagine that the winning numbers that have been extracted are 9, 19, 4, 46, 32, 21, with this exact order, then a person who selected 4, 32, 9, 46, 19, 21 qualifies as a winner or not? Of course, the answer depends on whether the ordering counts or not: if it does, then the person did not win, otherwise, the person won.

We realize that *in probability we have 4 different ways of counting*, which are visualized in the following table.

	Without Replacement	**With Replacement**
Ordered		
Unordered		

Table 1.1: Four different ways of counting.

In order to fill this table with the corresponding number of counting, we must introduce two definition that will help us in the following analysis.

Def. 1.3.1 (Permutation) *Given the set composed by the first n positive integers*

$$I_n := \{1, 2, \ldots, n\},$$

a permutation is defined to be a one-to-one and onto function on I_n.

In more concrete terms, *a permutation is simply an exchange of n labels such that labels cannot disappear or occupy more than one slot.*

The identity function $id : I_n \to I_n$, $id(n) = n$, is called the *trivial permutation* because it is the only one that leaves each label unaltered.

A non-trivial permutation of $I_4 = \{1, 2, 3, 4\}$ is, for example, $\{3, 1, 4, 2\}$. How many permutations of the n labels contained in I_n can exist? To answer this question we can reason as follows:

- we start occupying the first slot of the novel n-label set by choosing one number in I_n
- to occupy the second slot we have at disposal $n - 1$ labels
- we will arrive at the end, where the last slot can be occupied only by the last number remained in I_n.

By the fundamental theorem of counting we have that there exist $n \cdot (n-1) \cdot (n-2) \cdots 1$ permutations. This number has a special name.

Def. 1.3.2 (Factorial) *For a positive integer n, $n!$ (read **n factorial**) is the product of all of the positive integers less than or equal to n, i.e.*

$$n! := n(n-1) \cdots 1,$$

since multiplying by 1 does not change anything, this multiplication is, in general, bypassed.

The factorial can be extended to encompass 0 by defining

$$0! := 1.$$

So, there exist $n!$ permutations of a set of n elements. For example there are $5! = 5 \cdot 4 \cdot 3 \cdot 2 = 120$ ways to permutate 5 elements and $8! = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 = 40320$ ways to permutate 5 elements.

Looking at these numbers we have the intuition that the factorial grows fast... actually it grows very, very!!, fast, e.g. $10! = 3628800$ and $14! = 87178291200$.

A very useful property of the factorial is that:

$$n! = n \cdot (n-1)!$$

e.g., $5! = 5 \cdot 4! = 5 \cdot (4 \cdot 3 \cdot 2)$.

Example 1.3.2 (**4-digit entry code**) Imagine that a friend of yours invites you to dinner at his/her place and he/she tells you the 4-digit code that you have to type in order to open the door of the building where your friend lives.

When you arrive there you recall that the code contains the numbers 1,4,7,9, but you forgot the order. Our many different combinations do you have to test? Following our previous reasoning, there are $4! = 4 \cdot 3 \cdot 2 = 24$ different codes that you may enter before managing to open the building door.

Unless you have a formidable memory, you should write down each code in order to avoid repeating it by mistake. If we make a rough estimation of 10 seconds for both actions (typing and writing down the code), then, in the worst case scenario in which the correct code is the 24th, we have to spend $240s = 4$ minutes before opening the door.

If the entry code has 5 digits instead of 4, then, repeating the same reasoning, there are 120 different codes and the worst case scenario in this case is of 20 minutes. This gives an idea of how fast the time delay due to the factorial grows!

Finally, we remark that in this example we have supposed that we recall the digits appearing in the code, if we do not and we have to select them from the 10 digits $0, 1, \ldots, 9$, then the count is much larger, as we are going to see. ◇

We are now ready to analyze all the possible combinations of Table 1.1. We will continue using the example of the entry code as a beacon to concretely understand the different types of counting.

1. *Ordered counting, without replacement*: in this case the order of the code digit counts, but the code is supposed to be built from different digits. There are 10 ways to find out the first digit, then 9 for the second, then 8 and finally 7. The fundamental theorem of counting gives:

$$10 \cdot 9 \cdot 8 \cdot 7 = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdots 1}{6 \cdot 5 \cdots 1} = \frac{10!}{6!} = \frac{10!}{(10-4)!} = 5040.$$

With the same time estimation of the previous example, if we have forgotten the different digits that appear in the code, the worst case scenario to enter in the building grows to 14 hours!

In general, **the number of ways to choose r items from $n \geqslant r$ items without replacement, but considering order is**

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

2. *Ordered counting, with replacement*: in this case the order of the code digit counts, and the digit may repeat in the code. So, there are 10 ways to select each digit in the 4 cases, so the counting is

$$10 \cdot 10 \cdot 10 \cdot 10 = 10^4 = 10000.$$

So, if we have forgotten the code digits and they may repeat, the worst case scenario to enter in the building grows to over 27 hours!

In general, **the number of ways to choose r items from $n \geqslant r$ items with replacement and considering order is**

$$n^r.$$

3. *Unordered counting, without replacement*: we computed in the item 1. the counting without replacement when the ordering must be taken into account, so what we must do in this case is to divide out the redundant 4! orderings. So, the total counting in this case is

$$\frac{10!}{4!\,(10-4)!} = \frac{10!}{4!\,6!} = \frac{10 \cdot \cancel{9}^{\,3} \cdot \cancel{8} \cdot 7 \cdot \cancel{6!}}{\cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{6!}} = 210,$$

which is significantly smaller than 5040 found in item 1, in fact the worst time delay in this case is 35 minutes.

The expression obtained with this type of counting has a special name.

Def. 1.3.3 (Binomial coefficient) *For non-negative integers n and r such that $n \geqslant r$, we define the binomial coefficient $\binom{n}{r}$, read **n choose r**, as follows*

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

Notice that, in particular,

$$\binom{n}{0} = \frac{n!}{0!(n-0)!} = \frac{n!}{1 \cdot n!} = 1,$$

$$\binom{n}{1} = \frac{n!}{1!(n-1)!} = \frac{n(n-1)!}{(n-1)!} = n,$$

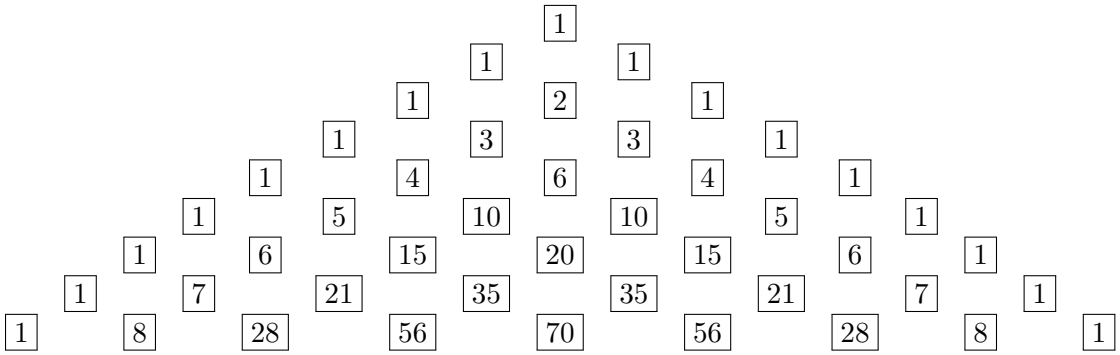
and

$$\binom{n}{n} = \frac{n!}{n!(n-n)!} = \frac{n!}{n!0!} = 1.$$

The reason for the name derives from the fact that the binomial coefficient represents the coefficients of the expansion of the positive integer power of a binomial:

$$(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r}.$$

These coefficients are exactly those appearing in the famous Pascal-Tartaglia triangle:



To resume, in general, **the number of ways to choose r items from $n \geqslant r$ items without replacement and not considering order is**

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

4. *Unordered counting, with replacement*: this is the most difficult case to count. At first glance, one could guess that the correct counting is obtained by dividing that of item 2., i.e. 10^4 by 4!, but this is incorrect (because it gives a smaller counting than the actual one). To explain the reason why in a meaningful way would take quite a long time, so we will only give the correct answer, which is

$$\binom{10+4-1}{4} = \binom{13}{4} = \frac{13 \cdot 12 \cdot 11 \cdot 10 \cdot \cancel{9!}}{4!\,\cancel{9!}} = \frac{13 \cdot \cancel{12} \cdot 11 \cdot \cancel{10}^{5}}{\cancel{12} \cdot \cancel{2}} = 715,$$

instead, $10^4/4 \simeq 417$. Under the circumstances of this item, the worst case scenario to open the building door would be about 2 hours.

In general, **the number of ways to choose r items from $n \geqslant r$ items with replacement and not considering order is**

$$\binom{n+r-1}{r}.$$

Table 1.2 summarizes the four cases just analyzed.

	Without Replacement	**With Replacement**
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

Table 1.2: Enumerating methods: number of ways to choose r items from $n \geqslant r$ items in the cases obtained by intersecting the row and the column.

1.3.1 Application of counting: enumerating outcomes

We can use the counting techniques learned in the previous section to calculate probabilities when the sample space S is **finite** and all the outcomes in S are **equally likely**.

To see how, suppose that $S = \{s_1, \ldots, s_n\}$ is a finite sample space. We call **cardinality** ('card') the number of the elements of a set, so in this case $\text{card}(S) = n$.

Now, saying that all the outcomes are equally likely means that the probability associated to each singleton event $\{s_i\}$, $i = 1, \ldots, n$, is identical. Since the probabilities must sum up to one, we must necessarily have

$$P(\{s_i\}) = \frac{1}{n},$$

for every outcome s_i, $i = 1, \ldots, n$.

Since every event E can be written as follows

$$E = \bigsqcup_{s_i \in E} \{s_i\},$$

using the axiom $K2$ we obtain

$$P(E) = P\left(\bigsqcup_{s_i \in E} \{s_i\}\right) \underset{K2}{=} \sum_{s_i \in E} P(\{s_i\}) = \sum_{s_i \in E} \frac{1}{n} = \frac{1}{n} \sum_{s_i \in E} 1 = \frac{\text{card}(E)}{n}. \tag{1.6}$$

Sometimes, the cardinality of the event E can be computed directly, some other times it requires the counting techniques that we have learned in the previous section. Let us see two examples, the first one in which the counting technique is direct and the second in which the more sophisticated enumeration methods that we have learned are necessary.

Example 1.3.3 (**Two dice**) Consider the rolling of two identical dice. The sample space S for this situation is of course given by all the **unordered couples**[2] of numbers $\{m, n\}$, with $m, n = 1, \ldots, 6$, hence $\text{card}(S) = 36$.

We want to compute the probability that we get *at least* one 6 with a single roll of the two dice. We can proceed like this:

- consider the event $E = \{\{6, m\},\ m = 1, \ldots, 6\}$
- and the event $F = \{\{m, 6\},\ m = 1, \ldots, 6\}$,

[2] it is custom to denote an **ordered couple with parenthesis** (,) and an **unordered couple with curly brackets** { , }.

Since we have specified that we are interested in the probability of having *at least* one 6, it is not important if 6 pops out after rolling the first or the second die. So, the probability that we are searching for is given by $P(E \cup F)$.

Recalling eq. (1.4), we have $P(E \cup F) = P(E) + P(F) - P(E \cap F)$, so we need to calculate the three probabilities on the right-hand side of this last equation.

To compute the probabilities of E and F we must take into account the fact that there are 6 ways in which both E and F may occur (because m runs through a set of 6 integer numbers), so eq. (1.6) implies

$$P(E) = P(F) = \frac{6}{36} = \frac{1}{6}.$$

Moreover,

$$E \cap F = \{\{6, 6\}\},$$

hence $\text{card}(E \cap F) = 1$ and eq. (1.6) implies that

$$P(E \cap F) = \frac{1}{36}.$$

Finally, the probability of having at least one 6 in the rolling of two identical dice is close to 1 chance over 3:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) = \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36} \simeq 0.3 = 30\%.$$

The number 6 has been selected just because, in many games, we aim at having the largest possible outcome, however it should be clear that there is nothing special about the number 6, so the probability calculated in this example holds true also for any other number between 1 and 6. $\diamond$

Example 1.3.4 (**Some probabilities at poker**)
Consider a standard poker deck of 52 cards, as depicted in Figure 1.6.

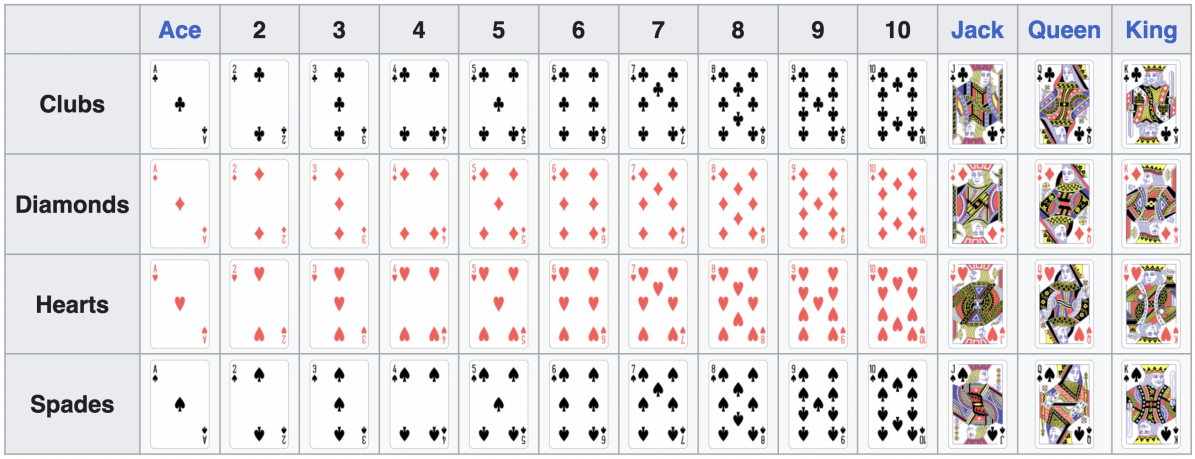


Figure 1.6: A standard deck of 52 cards. Source: Creative Commons.

The canonical terminology in poker is the following.

- **Hand**: 5 cards from a well shuffled deck
- **Kind**: the 4 kinds are Clubs, Diamonds, Hearts and Spades
- **Denomination**: the 13 denominations are the numbers associated to each card, from 1 (ace) to 13 (king).

Type of counting:

- In poker, cards are **always sampled without replacement** from the deck.

In this example we *choose* to look at the hand all at once, not minding when the 5 cards have been given to us. This clearly corresponds to an **unordered sampling**.

So, our probability counting in this example is **unordered and without replacement**. As a consequence, the sample space S of this example is

$$S = \{\text{All possible hands obtained from the deck, unordered and without replacement}\}.$$

The number of different unordered hands without replacement can be obtained from table 1.2

	Without Replacement	**With Replacement**
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

using $n = 52$ and $r = 5$, which gives

$$\binom{52}{5} = 2598960 = \text{card}(S).$$

Using formula (1.6), we can assign to the event

$$E = \{\text{having one hand of 5 specific cards}\} \subset S$$

the following probability

$$P(E) = \frac{1}{2598960} \simeq 4 \cdot 10^{-7} = 0.0000004\%$$

We see that the probability of guessing exactly the cards in one hand is very very small, less than 1 chance over 250000, as it was expected.

Now, let us compute the probability of having, for instance, a **poker of four aces** in our hand. We can answer this question by counting how many different hands there are with four aces and then using formula (1.6). If we specify that four of the cards are aces, then there are $52 - 4 = 48$ different ways of specifying the fifth card, i.e. 48 hands with fours aces in them. Thus formula (1.6) implies

$$P(\text{4 aces in one hand}) = \frac{48}{2598960} = \frac{1}{54145} \simeq 2 \cdot 10^{-5} = 0.00002\%,$$

so, there is less than 1 chance over 50000 to have a hand with a poker of four aces.

Let us now compute the probability of having at least **fours cards of the same kind** in our hand. The counting here can be decomposed in two tasks:

1. the first is to sample the cards of the same kind, but since there are 13 cards of the same kind, it follows that this task can be accomplished in 13 different ways

2. the second is to sample the fifth card from the remaining cards in the deck, which are $52 - 4 = 48$. Pay attention that we have demanded to have *at least* four cards of the same kind, thus we have not excluded the possibility of having five of them, this is why *all* the remaining cards of the deck must be counted and not only those having another kind with respect to the previous four.

The fundamental theorem of counting implies that the total number of hands with at least four cards of the same kind is $13 \cdot 48 = 624$, hence:

$$P(\text{at least 4 cards of same kind in a hand}) = \frac{624}{2598960} = \frac{1}{4165} \simeq 0.00024 = 0.024\%,$$

so, this time, we find a probability of less than 1 chance over 4000 to have a hand with at least for cards of the same kind, which is much likelier than having 4 aces.

Now let us calculate the probability that in our hand we have **exactly one pair of two kinds**, e.g. a 7 of spades and a 7 of hearts, but *not* two pairs, three of a (different) kind, etc.

This is a actually very interesting example because it will force us to use different counting techniques.

- In order to define the pair, we must specify a *denomination*: ace, two, three, etc. We know that number of the denominations is 13, from the ace to the king

- There are four *kinds* of cards, so we must also specify the two kinds among them. Since this specification is unordered and without replacement, from Table 1.2 we know that the number of choices that we can make is $\binom{4}{2} = \frac{4!}{2!\,2!} = 6$.

- The other three cards must have a different denomination than the one chosen for our pair, otherwise we will have three of a kind or more and we explicitly said that we want to exclude that possibility. Again, no ordering or replacement is involved in the choice of the remaining 12 denominations for the 3 cards that remain to compose our hand, so the counting for the specification of the these denominations is

$$\binom{12}{3} = \frac{12!}{3!\,9!} = \frac{12 \cdot 11 \cdot 10 \cdot 9!}{6 \cdot 9!} = 220.$$

- Finally, for each one of the three remaining cards, their kind can be selected among the four possible ones, which can be done in $4 \cdot 4 \cdot 4 = 4^3 = 64$ ways. ◇

By taking all the items into account, we have that the probability of having exactly one pair in our hand is

$$P(\text{having exactly one pair}) = \frac{13 \cdot 6 \cdot 220 \cdot 64}{2598960} = \frac{1098240}{2598960} \simeq 0.42 = 42\%,$$

which is by far the likeliest situation examined in this example.

We resume hereby what we have done in this example:

- the samples space S that we have considered consists in all the hands of 5 cards that we can get from a well-shuffled deck
- they come in a finite number n and they are equally likely
- so, given an event E associated with the game of poker, we can use eq. (1.6)

$$P(E) = \frac{\text{card}(E)}{n}$$

  to compute $P(E)$
- we have considered four types of events
  - E_1 := guessing exactly the 5 cards of one hand
  - E_2 := having 4 aces in one hand
  - E_3 := having at least 4 cards of same kind in one hand
  - E_4 := having exactly one pair
- using the counting methods that suit each event E_j, $j = 1, 2, 3, 4$, we have computed their probabilities obtaining
  - $P(E_1) \simeq 0.0000004\%$
  - $P(E_2) \simeq 0.00002\%$
  - $P(E_3) \simeq 0.024\%$
  - $P(E_4) \simeq 42\%$.

1.4 Conditional probability

All of the probabilities that we have dealt with so far have been *unconditional*: a sample space S was defined and all probabilities were calculated with respect to S. In many instances, however, we are in a position to **update the sample space based on new information**. In such cases, we want to be able to update probability calculations or, using the typical language of probability, to calculate 'conditional probabilities'. This name is chosen because the novel information acquired on an event may condition the probability that another event occurs, unless, as we will see, if the events are *independent*.

As always, the best way to introduce a new concept is through an example, which will also help us understanding the reason underlying the definition of conditional probability.

Example 1.4.1 (**Motivational example to define conditional probability**)
Four cards are dealt from the top of a well-shuffled deck. What is the probability that they are the four queens? We can answer this question using two methods.

1. *First method*. Let us use the counting techniques learned in the previous section. The number of distinct groups of four cards, unordered and without replacement, is

$$\binom{52}{4} = 270725.$$

   Only one of these groups consists of the four queens and every group is equally likely, so, thanks to eq. (1.6), the probability of being dealt all four queens is $1/\binom{52}{4}$.

2. *Second method*. Now we use an 'updating' argument: since there are four queens in the deck, the probability that the first card is a queen is $4/52$. *Given that the first card is a queen*, there remain three queens in 51 cards left, so the probability that the second card is a queen is $3/51$. Iterating this argument and using the fundamental theorem of counting, we get the desired probability as

$$\frac{4}{52}\cdot\frac{3}{51}\cdot\frac{2}{50}\cdot\frac{1}{49} = \frac{1}{270725} = \frac{1}{\binom{52}{4}}.$$

   This second method of solving the problem is closer to the concept of conditional probability because it shows very clearly that the act of taking out a queen from the deck conditions the number of remaining queens, hence the sample space must be updated! $\diamond$

If we want to formalize and generalize the lesson learned from this last example, we must isolate two essential features of the updating process:

- if an event occurred, then the conditional probability of this event must give back 1
- all further occurrences must be calibrated with respect to the novel sample space and not to the original one anymore. In particular, if an event is incompatible with the novel sample space, i.e. if it cannot occur simultaneously with the event already occurred, its conditional probability must be 0.

The definition of conditional probability that grasps these two essential features is the following.

Def. 1.4.1 (Conditional probability) *If E and F are events in a sample space S and $P(F) > 0$, then the conditional probability of E given F, written $P(E|F)$, is*

$$P(E|F) = \frac{P(E \cap F)}{P(F)}. \tag{1.7}$$

Let us verify that this definition is coherent with what we expect:

- if we compute the conditional probability of F given F we get

$$P(F|F) = \frac{P(F \cap F)}{P(F)} = \frac{P(F)}{P(F)} = 1,$$

  this means that S has been updated to the novel sample space F

- if E and F are incompatibles, i.e. $E \cap F = \varnothing$, then $P(E \cap F) = 0$ and so $P(E|F) = 0$.

Example 1.4.2 (Follow up of example 1.4.1)

In Example 1.4.1 we have seen that the probability of getting all four queens is small. Let us now see how the conditional probabilities change given that some queens have already been drawn, this will show one of the first *counter-intuitive facts of conditional probability.*

Four cards will again be dealt from a well-shuffled deck, and, alongside the event E considered in example 1.4.1, i.e. $E :=$ 4 queens in 4 cards, we define these other three events:

$$F_k = k \text{ queens in } k \text{ cards}, \ k = 1, 2, 3.$$

Notice that if the event E occurs, then the first card was a queen (1 queen in 1 card), the second too (2 queens in 2 cards) and the third too (3 queens in 3 cards), this means that if E occurs, then also F_k, $k = 1, 2, 3$, occur. So $E \subset F_k$, hence $E \cap F_k = E$, for all $k = 1, 2, 3$.

It follows that

$$P(4 \text{ queens in } 4 \text{ cards} \,|\, k \text{ queens in } k \text{ cards}) = P(E|F_k) = \frac{P(E \cap F_k)}{P(F_k)} = \frac{P(E)}{P(F_k)}.$$

We already computed $P(E)$ in the previous example:

$$P(E) = \frac{1}{\binom{52}{4}}.$$

Now let us compute $P(F_k)$:

- the number of possible distinct queens in a group of k cards without replacement and without considering the ordering is $\binom{4}{k}$
- the number of possible distinct sets of k cards in the deck without replacement and without considering the ordering is $\binom{52}{k}$,

so

$$P(F_k) = \frac{\binom{4}{k}}{\binom{52}{k}}.$$

Finally,

$$P(E|F_k) = \frac{P(E)}{P(F_k)} = \frac{1}{\binom{52}{4}} \frac{1}{\frac{\binom{4}{k}}{\binom{52}{k}}} = \frac{\binom{52}{k}}{\binom{52}{4}\binom{4}{k}},$$

using the definition of binomial coefficient we can simplify the fraction as follows

$$P(E|F_k) = \frac{52!\,4!\,(52-4)!\,k!\,(4-k)!}{k!\,(52-k)!\,52!\,4!} = \frac{48!\,(4-k)!}{(52-k)!} = \frac{1}{\binom{52-k}{4-k}}.$$

Letting k run over 1,2 and 3 we find the following conditional probabilities:

- $P(4 \text{ queens in 4 cards} \,|\, 1 \text{ queen in 1 card}) = 0.00005 = 0.005\%$
- $P(4 \text{ queens in 4 cards} \,|\, 2 \text{ queens in 2 cards}) \simeq 0.0008 = 0.08\%$
- $P(4 \text{ queens in 4 cards} \,|\, 3 \text{ queens in 3 cards}) \simeq 0.02 = 2\%$.

This of course contradicts our common sense: we do not expect the probability of being given a queen to increase given that we already had a queen in the previous occasion!

Common sense works well when we deal with unconditional probabilities, but **when we consider conditional probabilities, we must leave our common sense aside** and use mathematics to avoid mistakes. ◇

Another counter-intuitive example is provided by the following problem presented by **Martin Gardner** in the 'Mathematical Games' column of the *Scientific American* in **1959**.

Example 1.4.3 (**The three prisoners problem (1)**)

- Three prisoners, A, B, and C, have worked hard to show that they are ready to be restored into the society. The warden of the prison received a phone call from the state governor who tells him that he chose *at random* the prisoner that will be set free among A,B and C. He reveals his choice but forbids the warden to reveal it to the prisoners before he arrives to the prison in person.
- The next day, A tries to get the warden to tell him who had been pardoned, but the warden refuses saying that he received precise instructions not to reveal that information.
- The day after that, A asks to be allowed to call by the phone the warden and asks him to be allowed to know at least the identity of one of the two prisoners who are going to remain in jail. Precisely, he tells the warden: '*If B is to be pardoned, give me C's name. If C is to be pardoned, give me B's name. And if I'm to be pardoned, flip a coin to decide whether to name B or C*'.
- The warden, *who knows probability*, tells A that he will call him back in an hour. After that time, he reveals to A that B will remain in prison.
- A, *who does not know probability*, uses just his common sense and thinks that his probability to get out of prison has increased from 1/3 to 1/2, because now the choice is between him and C.

- A decides to secretly tell C that B will remain in prison. C, *who knows probability*, is happy to hear that because he knows that his probability of being pardoned has increased from 1/3 to 2/3, but he does not say a word to A... :)

Let a, b, and c **denote the events that A, B, or C is pardoned**, respectively, *before* the warden says anything.

Since the governor chose at random who to pardon, before the warden says anything we have identical probabilities:

$$P(a) = P(b) = P(c) = \frac{1}{3}.$$

Now, let w **denote the event that the warden tells A that B remains in prison**. Since this event indeed occurs, the previous probabilities must be updated to conditional ones. Those of interest for us are:

$$P(a|w) = \frac{P(a \cap w)}{P(w)} \quad \text{and} \quad P(c|w) = \frac{P(c \cap w)}{P(w)}.$$

Let us analyze the different probabilities that appear above:

- $P(w)$ is the probability that the warden tells A that B remains in prison, but this probability is 1/2 because the warden will always name either B or C! So:

$$P(w) = \frac{1}{2}.$$

- $P(a \cap w)$ is the probability that A is pardoned *and* the warden tells him that B remains in prison. Recall that if A is pardoned, the warden agreed to flip a coin and decide to tell him that either B or C will remain in prison. Crucially, the act of flipping a coin is associated to a probability 1/2, hence $P(a \cap w) = P(a) \cdot 1/2$, i.e.

$$P(a \cap w) = \frac{1}{6}.$$

- $P(c \cap w)$ is the probability that C is pardoned *and* the warden tells A that B remains in prison. In this case the behavior of the warden is deterministic and not probabilistic! In fact, if C is pardoned, then the warden agreed to tell A that B will remain in prison, and there is nothing probabilistic in that action. For this reason:

$$P(c \cap w) = P(c) = \frac{1}{3}.$$

In conclusion:

$$P(a|w) = \frac{1/6}{1/2} = \frac{1}{3},$$

instead

$$P(c|w) = \frac{1/3}{1/2} = \frac{2}{3}.$$

We see that **conditional probabilities can be quite slippery and require a careful interpretation**. A tool that help us correcting our wrong intuition about conditional probabilities is the Bayes theorem, that allows us updating probabilities by incorporating new information and will be the topic of the next section. ◇

1.5 The Bayes theorem and statistical independence

The importance of Bayes's theorem cannot be overestimated, to the point that the mathematician and physicist Harold **Jeffreys** wrote in a 1973 book that '***Bayes' theorem is to the theory of probability what the Pythagorean theorem is to geometry***'.

It is quite remarkable that one of the most important laws of probability can be obtained by a very simple rearrangement of the equation that expresses conditional probability, i.e. formula (1.7).

First of all we can write it as follows

$$P(E \cap F) = P(E|F)P(F), \tag{1.8}$$

which gives a useful formula for calculating probabilities of simultaneous events. Given that $E \cap F = F \cap E$, we can exchange the events E and F on the right-hand side, thus obtaining

$$P(F \cap E) = P(F|E)P(E). \tag{1.9}$$

Equating the right-hand sides of eqs. (1.8), (1.9) gives $P(E|F)P(F) = P(F|E)P(E)$, which is usually written in the form contained in the following theorem.

Theorem 1.5.1 (Bayes' theorem) *If E and F are two events of the sample space S, then it holds that*

$$P(F|E) = P(E|F)\frac{P(F)}{P(E)}, \qquad P(E) > 0. \tag{1.10}$$

Eq. (1.10) shows that **in general, conditional probability is not symmetric**, in fact it is symmetric if and only if $P(E) = P(F)$. So, Bayes' theorem says that **conditional probabilities can be 'turned around' up to a factor**, given by the fraction of the unconditional probabilities.

The discoverer of this formula is Thomas **Bayes** (1701-1761) who never expressed it in mathematical terms, but explained it in words.

He did not consider it an important result, so he never bothered publishing it. However, after his death, his family asked his friend and mathematician Richard **Price** to look for possible interesting results in Bayes' papers. Price recognized the importance of Bayes' ideas and published his theorem posthumously in **1763**. The great French mathematician Pierre Simon de **Laplace** (1749-1827) independently found Bayes's theorem, unaware of Bayes and Price work, and published his version in 1774.

The probabilities appearing in formula (1.10) have special names:

- $P(F)$: **prior probability** of the event F
- $P(E)$: **marginal probability**, it serves as a normalization factor
- $P(E|F)$: **likelihood**
- $P(F|E)$: **posterior probability** of F given that E occurs.

The Bayes theorem can be generalized thanks to the following result.

From theorem 1.2.4, recall that if $\{F_j\}_{j=1}^n$ is a partition of the sample space S, then it holds that

$$P(E) = \sum_{j=1}^{n} P(E \cap F_j) = P(E \cap F_1) + \cdots + P(E \cap F_n). \tag{1.11}$$

This formula is often called **law of total probability** and, since for all event F we have $S = F \sqcup F^c$, it implies

$$P(E) = P(E \cap F) + P(E \cap F^c). \tag{1.12}$$

Corollary 1.5.1 (Generalized Bayes' theorem) *The Bayes theorem, eq.* (1.10)*, can be extended to any partition* $\{F_j\}_{j=1}^n$ *of the sample space* S *as follows*

$$P(F_j|E) = \frac{P(E|F_j)P(F_j)}{\sum_{j=1}^{n} P(E|F_j)P(F_j)}. \tag{1.13}$$

In particular, for the partition $S = F \sqcup F^c$ *we have*

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|F^c)P(F^c)}. \tag{1.14}$$

Proof. Applying Bayes' theorem to each set of the partition we have

$$P(F_j|E) = P(E|F_j)\frac{P(F_j)}{P(E)}, \quad j = 1, \ldots, n, \tag{1.15}$$

but, thanks to eq. (1.11), the marginal probability can be written as

$$P(E) = \sum_{j=1}^{n} P(E \cap F_j),$$

moreover, by definition of conditional probability,

$$P(E|F_j) = \frac{P(E \cap F_j)}{P(F_j)} \quad \text{so} \quad P(E \cap F_j) = P(E|F_j)P(F_j), \quad j = 1, \ldots, n,$$

hence

$$P(E) = \sum_{j=1}^{n} P(E|F_j)P(F_j).$$

Finally, introducing this expression of the marginal probability in eq. (1.15) we find

$$P(F_j|E) = \frac{P(E|F_j)P(F_j)}{\sum_{j=1}^{n} P(E|F_j)P(F_j)}.$$

□

Let us re-elaborate the three prisoners problem in view of the Bayes theorem.

Example 1.5.1 (**The three prisoners problem (2)**) We recall that in Example 1.4.3 we indicated with a, b, c the event that the prisoner A, B, C, respectively, will be pardoned, and with w the event that the warden tells A that prisoner B remains in jail. From Example 1.4.3 we know that $P(a) = P(b) = P(c) = 1/3$.

We now remark that the events a, b, c are disjoint, because only one prisoner is pardoned, and their union gives the sample space of the problem 'one prisoner will be pardoned and two will remain in jail'.

Hence, a, b, c is a partition and we can use the generalized Bayes' theorem to compute the posterior probability of A and C being pardoned:

$$P(a|w) = \frac{P(w|a)P(a)}{P(w|a)P(a) + P(w|b)P(b) + P(w|c)P(c)},$$

$$P(c|w) = \frac{P(w|c)P(c)}{P(w|a)P(a) + P(w|b)P(b) + P(w|c)P(c)}.$$

Let us analyze the different cases:

- if A will be pardoned, the warden can tell A that either B or C will remain in jail, hence $P(w|a) = \frac{1}{2}$
- if B will be pardoned, the warden cannot tell A that B will remain in jail, so $P(w|b) = 0$
- if C will be pardoned, the warden can only tell A that B remains in jail, so $P(w|c) = 1$.

Introducing these values in the previous equations we find

$$P(a|w) = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{3}.$$

$$P(c|w) = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{2}{3}.$$

confirming the analysis of Example 1.4.3, but with a more straightforward argument. ◇

The three prisoners problem is mathematically equivalent to another very famous apparent paradox, called the Monty Hall problem.

Example 1.5.2 (**The Monty Hall problem**) This problem was posed in the form of a probability puzzle on the American television game show '*Let's Make a Deal*' and named after its original host, *Monty Hall*. It was then formalized as follows in 1990:

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door and the host, who knows what's behind the doors, opens another door which has a goat. He then says to you, 'Do you want to switch your choice?'

Probabilistically speaking, is it to your advantage to switch your choice or not?

We first analyze the problem without math and then we formalize it with the Bayes theorem to confirm our analysis.

- Suppose you do not change our initial choice, then there is only one possibility to win the car, i.e. when the car is behind the door you chose. Since there are three doors, the probability of winning the car is 1/3.
- Suppose now that you change our initial choice. To fix the ideas, suppose that the car is behind door 1, then we must subdivide the analysis in three sub-cases:
  1. you chose door 1 then, independently of the door that the host opens, if you switch door you lose
  2. you chose door 2, then the host opens door 3 revealing the goat, if you switch to door 1 you win the car
  3. you chose door 3, then the host opens door 2 revealing the goat, if you switch to door 1 you win the car.

Hence, if you remain stuck with your initial choice, you have probability 1/3 to win the car, instead, if you modify your initial choice, you have 2 possibilities over 3, i.e. a probability of 2/3, to win.

Let us now formalize the problem mathematically using Bayes's theorem, which is the appropriate tool because *it describes the change in probability of a given event based on new information*. In this case, we had an initial guess, but the host showed another door with a goat behind it, which added further knowledge about what is behind the three doors.

To fix the ideas, suppose our initial guess was door number 1, and the host revealed that behind door number 3 there is a goat, as shown in Figure 1.7.

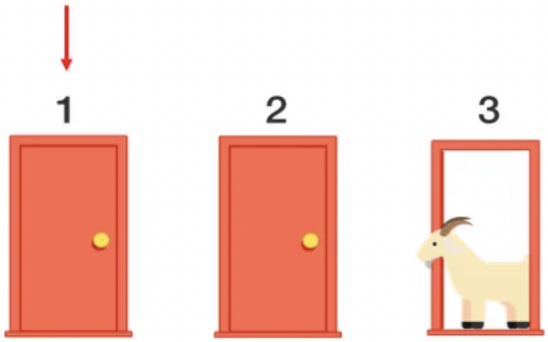


Figure 1.7: A case study in the Monty Hall problem.

Let us denote with C the event '*the car is behind door 1*' and with M the event '*Monty has revealed a goat behind a door different than 1*'.

By eq. (1.14), the *posterior* probability that the car is behind door 1 given that Monty opens door 3 revealing a goat is

$$P(C|M) = \frac{P(M|C)P(C)}{P(M|C)P(C) + P(M|C^c)P(C^c)},$$

where C^c is the event '*a goat is behind door 1*'.

We have:

- $P(C)$ is the prior probability that the car is behind door 1, without knowing about the door that the host reveals. This is of course 1/3 because at the beginning each door has the same probability of hiding the car
- $P(M|C)$ is the following probability: given C, i.e. that there is *a car* behind door 1, Monty opens a door different than 1 with a goat behind it. Since Monty always shows a goat, this probability is equal to 1
- $P(M|C^c)$ is the following probability: given C^c, i.e. that there is *a goat* behind door 1, Monty opens a door different than 1 with a goat behind it. Again, since Monty always shows a goat, this probability is equal to 1
- $P(C^c)$ is the prior probability that we did not pick the door with the car behind it. Since the door that we chose has either the car behind it or not we have $P(C) + P(C^c) = 1$, it follows that $P(C^c) = 1 - 1/3 = 2/3$.

So, we have

$$P(C|M) = \frac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3}} = \frac{1}{3},$$

i.e., the probability that the car is behind door 1 is completely unchanged by the evidence shown by Monty and it has not increased to 1/2 as our intuition would incorrectly imply! (Notice the analogy with the three prisoners problems...)

After this computation, the situation is as follows:

- the car can be behind door 1, with probability 1/3
- or it can be behind the door that the host did not open, with a probability x
- but, of course, it cannot be behind the door that the host opened, because it revealed a goat! Hence, this probability is 0.

Since these three probabilities must sum up to 1 we have

$$\frac{1}{3} + x + 0 = 1 \iff x = 1 - \frac{1}{3} = \frac{2}{3},$$

therefore, switching gives twice the probability to find the car behind the door.

However, it is very important to underline that this does not mean that we will surely win the car if we change our mind! **A higher probability is not at all a synonym of certain success**. ◇

1.5.1 Statistical independence

Both the three prisoners and the Monty Hall problems show that, in some cases, the occurrence of a particular event has no effect on the probability of another event: the fact that the warden tells prisoner 1 that either one of the other prisoner will remain in jail does not modify his chances of being pardoned and, analogously, the fact that the Monty opens a door revealing a goat has no influence on the probability on winning the car if we stuck with our initial choice.

Using the notation of conditional probability, if, given F, the probability of E remains the same, then

$$P(E|F) = P(E). \tag{1.16}$$

If this the case, then the Bayes theorem (1.10) becomes

$$P(F|E) = P(E|F)\frac{P(F)}{P(E)} \underset{(1.16)}{=} P(E)\frac{P(F)}{P(E)} = P(F) \iff P(F|E) = P(F), \tag{1.17}$$

i.e., we have just discovered that *the lack of influence is symmetric*: the occurrence of E has no effect on the probability of the occurrence of F.

Considering the definition of conditional probability

$$P(F|E) = \frac{P(E \cap F)}{P(E)} \iff P(E \cap F) = P(E)P(F|E),$$

it follows that, if eq. (1.17) holds, i.e. if the events E and F do not influence each other, then

$$P(E \cap F) = P(E)P(F),$$

this formula happens to grasp the essence of mutual independence between events.

Def. 1.5.1 *Two events, E and F, are* ***statistically independent*** *if*

$$P(E \cap F) = P(E)P(F).$$

Example 1.5.3 (**Chevalier de Méré**) The gambler introduced at the beginning of the chapter, Chevalier de Méré, who involuntarily started the theory of probability, was particularly interested in the event that he could throw at least 1 six in 4 rolls of a die. We have

$$P(\text{at least 1 six in 4 rolls of a die}) = 1 - P(\text{no six in 4 rolls})$$

but rolls of a die are independent events, so

$$P(\text{at least 1 six in 4 rolls of a die}) = 1 - \prod_{j=1}^{4} P(\text{no six in roll } j),$$

where the symbol $\prod$ is the analogous of $\sum$ but for products. On any roll, the probability of *not* trowing a six is 5/6, so

$$P(\text{at least 1 six in 4 rolls of a die}) = 1 - \left(\frac{5}{6}\right)^4 \simeq 0.52,$$

so there is a 52% chance to throw at least a six with 4 rolls of a die. ◇

Definition 1.5.1 is preferred to eqs. (1.16) and (1.17) because it treats intrinsically E and F symmetrically (in fact, $E \cap F = F \cap E$) and this permits a simpler generalization of the concept of independence to more than two events, as we are now going to discuss.

We might think events E, F, and G are independent if $P(E \cap F \cap G) = P(E)P(F)P(G)$. However, this is not the correct condition, as the following example shows.

Example 1.5.4 Let us consider the tossing of two dice. This time we define the sample space to be the set of the 36 ordered pairs formed from the numbers 1 to 6:

$$S = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(2,6),\ldots,(6,1),\ldots,(6,6)\}.$$

Define the following events:

$$E = \{\text{doubles appear}\} = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\},$$

$$F = \{\text{the sum is between 7 and 10}\},$$

$$G = \{\text{the sum is 2 or 7 or 8}\}.$$

A direct counting among the 36 possible outcomes in S (you are encouraged to do it) gives that there are 6 outcomes for E, 12 for F and 18 for G, so, thanks to eq. (1.6), the probabilities of these events are

$$P(E) = \frac{1}{6}, \quad P(F) = \frac{1}{3}, \quad P(G) = \frac{1}{2}.$$

Now, the only possible way to obtain the simultaneous occurrence of these events, i.e. their intersection $E \cap F \cap G$, is when the sum of the dice is 8 with two 4s. In fact, the sum of the occurrences in E is always an even number, but the sum 2 does not belong to F and the sum 10 does not belong to G, so only the sum of 8 with two 4s (in order to belong to E) can belong to the intersection of the three sets. So,

$$P(E \cap F \cap G) = P(\text{the sum is 8, composed of double 4s}) = \frac{1}{36} = \frac{1}{6}\cdot\frac{1}{3}\cdot\frac{1}{2} = P(E)P(F)P(G).$$

However, $F \cap G$: sum equals 7 or 8 and this event can occur is 11 ways (check it...), hence

$$P(F \cap G) = P(\text{sum equals 7 or 8}) = \frac{11}{36} \neq P(F)P(G).$$

Similarly, you can verify as a very useful exercise that

$$P(E \cap F) \neq P(E)P(F).$$

Therefore, **the requirement $P(E \cap F \cap G) = P(E)P(F)P(G)$ is not a condition strong enough to guarantee pairwise independence!** $\diamond$

In light of the previous example, a second attempt to define independence for three events might be to define E, F, and G to be independent if all the pairs are independent. However, unfortunately, also this condition fails, as shown in the following example.

Example 1.5.5 This is *not an explicit example, but an abstract* one which, however, may describe a particular situation that may concretely happen.

Imagine the sample space S consists of the $3! = 6$ permutations of the letters a, b, and c along with the three triples of each letter. Thus:

$$S = \{aaa,\ bbb,\ ccc,\ abc,\ bca,\ cba,\ acb,\ bac,\ cab\}.$$

Furthermore, suppose that each outcome of S has the same probability of $1/9$. Define the event

$$E_i = \{i\text{-th place in the triple is occupied by } a\},$$

then

$$E_1 = \{aaa,\ abc,\ acb\}, \quad E_2 = \{aaa,\ bac,\ cab\}, \quad E_3 = \{aaa,\ bca,\ cba\}.$$

So

$$P(E_1) = P(E_2) = P(E_3) = \frac{3}{9} = \frac{1}{3}.$$

Moreover,

$$E_1 \cap E_2 = E_1 \cap E_3 = E_2 \cap E_3 = \{aaa\},$$

hence

$$P(E_1 \cap E_2) = P(E_1 \cap E_3) = P(E_2 \cap E_3) = \frac{1}{9}.$$

Clearly

$$P(E_i \cap E_j) = \frac{1}{9} = \frac{1}{3} \cdot \frac{1}{3} = P(E_i)P(E_j), \quad i, j = 1, 2, 3,\ i \neq j,$$

so the events E_1, E_2, E_3 are pairwise independent. However, it also holds that

$$E_1 \cap E_2 \cap E_3 = \{aaa\},$$

but

$$P(E_1 \cap E_2 \cap E_3) = \frac{1}{9} \neq \frac{1}{27} = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} = P(E_1)P(E_2)P(E_3).$$

So, the pairwise independence is not guarantee to imply $P(E_1 \cap E_2 \cap E_3) = P(E_1)P(E_2)P(E_3)$ and, as we have see in the previous example, this last formula is not guarantee to imply pairwise independence. ◇

The previous two examples show that simultaneous (or mutual) independence of a collection of more than two events requires *an extremely strong condition*. The right one is the request that the probability of the product of every possible intersection between events must be equal to the product of the probabilities of the single events.

To fix the ideas, if we have 4 events E_1, E_2, E_3, E_4, then they are independent if

$$P(E_1 \cap E_2 \cap E_3 \cap E_4) = P(E_1)P(E_2)P(E_3)P(E_4), \quad P(E_1 \cap E_4) = P(E_1)P(E_4),$$

$$P(E_2 \cap E_3 \cap E_4) = P(E_2)P(E_3)P(E_4), \quad \text{etc} \ldots \text{until exhaustion of all cases!}$$

Def. 1.5.2 *The events $E_1, \ldots, E_n$ are* ***mutually independent*** *if,* ***for any*** *collection $(E_{i_j})_{j=1}^{k}$, $1 \leqslant i_j \leqslant n$, $j = 1, \ldots, k \leqslant n$, we have*

$$P\left(\bigcap_{j=1}^{k} E_{i_j}\right) = \prod_{j=1}^{k} P(E_{i_j}).$$

Example 1.5.6 (**Three tosses of a coin**) Consider the experiment of tossing a coin three times. An outcome of this experiment must indicate the result of each 3-times toss. For example, HHT indicates that two heads and then tails were observed. The sample space for this experiment has 8 outcomes, namely

$$S = \{HHH,\ HHT,\ HTH,\ THH,\ TTH,\ THT,\ HTT,\ TTT\}.$$

Let H_j, $j = 1, 2, 3$, denote the event in which the j-th toss is heads:

$$H_1 = \{HHH,\ HHT,\ HTH,\ HTT\},$$

$$H_2 = \{HHH,\ HHT,\ THH,\ THT\},$$

$$H_3 = \{HHH,\ HTH,\ THH,\ TTH\}.$$

If the coin is fair, then we must assign probability $1/8$ to each outcome and since all the events contain 4 outcomes we find:

$$P(H_1) = P(H_2) = P(H_3) = \frac{4}{8} = \frac{1}{2},$$

so that the coin has equal probability of landing heads or tails on each toss.

Let us verify that the events H_1, H_2, and H_3 are mutually independent. We start by checking independence for the intersection of the three events:

$$P(H_1 \cap H_2 \cap H_3) = P(\{HHH\}) = \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(H_1)P(H_2)P(H_3).$$

Now we check the independence of each pair:

$$P(H_1 \cap H_2) = P(\{HHH,\ HHT\}) = \frac{2}{8} = \frac{1}{4} = P(H_1)P(H_2),$$

$$P(H_1 \cap H_3) = P(\{HHH,\ HTH\}) = \frac{2}{8} = \frac{1}{4} = P(H_1)P(H_3),$$

$$P(H_2 \cap H_3) = P(\{HHH,\ THH\}) = \frac{2}{8} = \frac{1}{4} = P(H_2)P(H_3).$$

Thus, H_1, H_2 and H_3 are mutually independent, i.e., the occurrence of heads in one toss has absolutely no effect on the occurrence of heads in the other tosses.

$\diamond$

Trivia: at the beginning of the 1990 Tom Stoppard's movie '*Rosencrantz and Guildenstern are dead*', Rosencrantz finds a coin and then toss it 92 times obtaining always heads... this causes Guildenstern to conclude that something is wrong with reality... would you guess the same after what you learned about probability?

We end this chapter with the answer to a natural question: does the independence of events E and F imply the independence of their complementary sets? The answer is positive.

Theorem 1.5.2 *If E and F are independent events, then the following pairs are also independent:*

1. *E and F^c*
2. *E^c and F*
3. *E^c and F^c.*

Proof.

1. To prove the first property, we must show that $P(E \cap F^c) = P(E)P(F^c)$. From property 1. of theorem 1.2.3 we have

   $$P(E \cap F^c) = P(E) - P(E \cap F) \underset{E,F \text{ ind.}}{=} P(E) - P(E)P(F) = P(E)(1 - P(F)) = P(E)P(F^c).$$

2. It follows immediately from 1. simply by interchanging the role of E and F.

3. To prove the third property, we must show that

   $$P(E^c \cap F^c) = P(E^c)P(F^c) = (1 - P(E))(1 - P(F)) = 1 - P(E) - P(F) + P(E)P(F).$$

   By using again property 1. of theorem 1.2.3 we have

   $$P(E^c \cap F^c) = P(E^c) - P(E^c \cap F) \underset{2.}{=} 1 - P(E) - P(E^c)P(F),$$

   having used in the second equality the fact that, thanks to property 2. of this theorem, if E and F are independent, then also E^c and F are. So,

   $$P(E^c \cap F^c) = 1 - P(E) - P(E^c)P(F) = 1 - P(E) - (1 - P(E))P(F) = 1 - P(E) - P(F) + P(E)P(F),$$

   as we wanted to verify.

□

1.6 Odds of a binary event and Bayes ratio

In this section we deal with binary events, that we define as follows.

Def. 1.6.1 *A binary event refers to an event that can have only two possible outcomes.*

Examples of outcomes of a binary events are: success/failure, true/false, yes/no, 1/0.

Examples of binary events are:

- *coin toss*, because the outcome can only be either heads or tails
- *pass/fail test*: a student either passes or fails an exam
- *light switch*: which can be on (1) or off (0).

A typical example of non-binary event is rolling a die because there are six possible outcomes: $1, 2, 3, 4, 5, 6$.

For binary events we can define the concept of odds.

Def. 1.6.2 (Odds ratio of a binary event) *Given a binary event E, its odds ratio is the fraction between the probability of occurrence and that of non-occurrence, i.e.*

$$o(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}.$$

- If $o(E) > 1$, then the event E is more likely to occur, and we say that **the odds are $o(E)$ to 1 in favor of the event E to occur**.
- If $o(E) < 1$, then the event E is more likely *not* to occur, and we say that **the odds are $o(E)$ to 1 in favor of the event E *not* to occur**.
- If $o(E) = 1$, we say that the odds of E are fifty-fifty.

This is the typical language used by *bookmakers*.

Let us reformulate Bayes' theorem in terms of odds. Consider another event F which may affect the event E. By Bayes theorem we know that

$$P(E|F) = P(F|E)\frac{P(E)}{P(F)} \quad \text{and} \quad P(E^c|F) = P(F|E^c)\frac{P(E^c)}{P(F)}. \tag{1.18}$$

The odds ratio $o(E)$ is updated as follows after the occurrence of F:

$$o(E|F) := \frac{P(E|F)}{P(E^c|F)} \underset{(1.18)}{=} P(F|E)\frac{P(E)}{\cancel{P(F)}} \cdot \frac{1}{P(F|E^c)}\frac{\cancel{P(F)}}{P(E^c)} = \frac{P(E)}{P(E^c)}\frac{P(F|E)}{P(F|E^c)}$$

but $P(E)/P(E^c) = o(E)$, so

$$o(E|F) = o(E) \cdot \frac{P(F|E)}{P(F|E^c)},$$

so the original odds ratio is updated by a coefficient called **Bayes factor** and expressed by

$$\mathscr{B} = \frac{P(F|E)}{P(F|E^c)}.$$

The higher the Bayes factor, the larger the odds ratio in favor of E after the event F and vice-versa.

Example 1.6.1 (**A left-handed criminal**)

A police investigation has determined that the perpetrator of a criminal act is left-handed. There is a person suspected for the crime who is indeed left-handed. Knowing that left-handed people are 10% of the population, determine by how much the ratio favorable to the suspect's guilt increases.

Let us define two events:

- E: '*the suspect is guilty*', which is a binary event
- F: '*the criminal is left-handed*'.

The conditional probability $P(F|E)$ is

$$P(F|E) = 1,$$

in fact, if E occurs, i.e. the suspect is guilty, then the suspect becomes the criminal and since the suspect is left-handed, the criminal is left-handed.

Instead, the conditional probability $P(F|E^c)$ is

$$P(F|E^c) = 0.1,$$

because E^c means that the suspect is not guilty, so the criminal is still unknown and we have the just the information that the criminal is left-handed, whose probability is 0.1=10%.

It follows that the Bayes factor is

$$\mathscr{B} = \frac{P(F|E)}{P(F|E^c)} = \frac{1}{0.1} = 10.$$

So, discovering a relatively rare feature in common between the criminal and the suspect increases a lot the probability that the suspect is indeed the criminal. Of course, the rarer the feature, the higher the rise in probability.

$\diamond$

Example 1.6.2 (**The rare disease problem**) After running a test, a person tested positive to a very rare disease which affects the 0.1% of the world population. The person asks the doctor about the accuracy of the test and it turns out that it correctly identifies 99% of people who have that disease.

1. What is the probability that the person actually has the rare disease?
2. Update the probability in the case that the person repeats the test in another laboratory and turns out positive again.

The answer to the first question is given by the conditional probability that the person actually has the disease after being tested positive, so we need to define the following events.

- H: hypothesis that the person actually has the disease. This event occurs when the hypothesis is true.
- E: event that the person tested positive.

By the generalized Bayes's theorem, the conditional probability of the hypothesis being true given that the event E occurs, i.e. probability to actually have the disease given that the test turned out to be positive, is

$$P(H|E) = \frac{P(E|H)P(H)}{P(H)P(E|H) + P(H^c)P(E|H^c)},$$

where

- $P(H)$: prior probability that hypothesis is true, i.e. that the person had the disease *before* the positive test result
- $P(H^c)$: probability that the person does not have the disease
- $P(E)$: probability of testing positive
- $P(E|H)$: conditional probability to test positive given that the person actually has the disease.
- $P(E|H^c)$: probability to test positive given that the person does not have the disease. This case is said to define a **false positive**.

The prior probability $P(H)$ is in general the most difficult thing to find out when applying Bayes's theorem and several times is no more than an educated guess. However, in this particular case a reasonable approximation of $P(H)$ is the probability of the disease in the population, which is $0.1\% = 0.001$. Another information that we know is the test accuracy, i.e. $P(E|H) = 99\% = 0.99$. So:

$$P(H|E) = \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + (1 - 0.99)(1 - 0.001)} \simeq 0.09 = 9\%$$

Despite this result seems very counter-intuitive because of the high accuracy of the test, if we think about the computation from a different angle, we can realize that this result is actually to be expected. Consider a population of 1000 people, then the probability that a person is affected by the disease is the 0.1% of 1000, which is exactly 1 person.

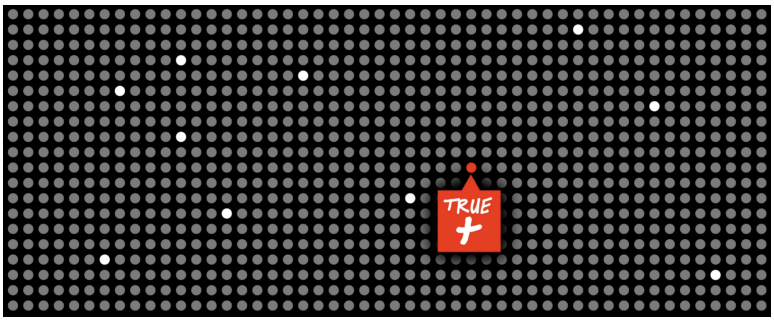


Figure 1.8: The test applied on a population of 1000 people, revealing a true person with disease and 10 false positives. Picture: courtesy of Derek Muller.

Suppose this is indeed the case, i.e. that one person is ill and we remain with 999 people. Since the test correctly identifies 99% of cases, there remains a 1% of false positives for 999 people, i.e. 10 false positives, as shown in Figure 1.8.

So the situation is the following: we have 11 positive tests, with 1 true ill person and 10 healthy people, hence the probability of being ill after being tested positive is $1/11 \simeq 0.09 = 9\%$, as our previous calculation showed.

This is the typical case that shows that *an accurate test is not necessarily a good predictive one*! In particular, for very rare diseases.

To overcome this problem, laboratories share data and perform crossed tests. Let us see why. Suppose that the person who actually has the disease undergoes another test in a different laboratory and turns out to be positive again. This time, we must update the old prior probability $P(H) = 0.001$ with the posterior probability $P(H|E) = 0.09$ after the result of the first test, which becomes the new prior. This time we have:

$$P_{\text{updated}}(H|E) = \frac{0.99 \cdot 0.09}{0.99 \cdot 0.09 + (1 - 0.99)(1 - 0.09)} \simeq \frac{0.0891}{0.0891 + 0.0091} \simeq 0.91 = 91\%$$

We see a dramatic increasing in the probability, which explains why (when it is possible) tests in different laboratories must be run to reduce the cases of false positives in rare diseases. $\diamond$

Example 1.6.3 (**The cookies problem**) Alice and Bob have both received as a gift a box of cookies:

1. in Alice's box there are 15 chocolate chip cookies and 5 vanilla ones
2. in Bob's box there are 8 chocolate chip cookies and 12 lemon curd ones.

By mistake, the boxes are mixed and one of their friends eats a chocolate chip cookie. What is the probability that this one belongs to Alice's box?

Of course we expect to find a probability larger than 50% in favor of Alice's box because of the presence of more chocolate chip cookies, but to quantify this value we need to use a Bayesian analysis.

Let us identify the elements of the problem:

- the sample space is $S = \{40 \text{ cookies}\}$
- E : event that a cookie belongs to Alice's box
- F : event that a cookie belongs to Bob's box
- C : event that a cookie in one box is a chocolate chip one.

By Bayes theorem, the probability that the friend ate a chocolate chip cookie which came from Alice's box is:

$$P(E|C) = P(C|E)\frac{P(E)}{P(C)}. \tag{1.19}$$

The prior probability that a cookie belongs to Alice's box or to Bob's box is identical:

$$P(E) = P(F) = \frac{1}{2}.$$

Since there is a total of $15 + 8 = 23$ chocolate chip cookies, the probability to eat one of them is

$$P(C) = \frac{23}{40}.$$

Now let us introduce the conditional probabilities: since there are 15 chocolate chip cookies in Alice's box, we have

$$P(C|E) = \frac{15}{20} = \frac{3}{4}.$$

Introducing the values just found in eq. (1.19) we find:

$$P(E|C) = P(C|E)\frac{P(E)}{P(C)} = \frac{3}{4}\frac{\frac{1}{2}}{\frac{23}{40}} = \frac{\frac{3}{2}}{\frac{23}{10}} = \frac{3}{2}\cdot\frac{10}{23} = \frac{15}{23} \simeq 0.65 = 65\%$$

$\diamond$

Example 1.6.4 (**The kittens problem**) A female cat just gave birth to two kittens and one of them is a male. What is the probability that also the other one is a male given that the average probability to give birth to a male among cats is 50%?

Our intuition leads us to think that there is 50% chance that the other kitten is a male. But we are dealing with a conditional probability, so we have to carefully formulate and solve the problem using the Bayesian analysis.

First of all, we must identify the sample space of the problem, which is given by the 4 possible couples of female-male kittens:

$$S = \{(F, F), (F, M), (M, F), (M, M)\}.$$

Then, if we define the event

$$E : \text{'at least one of the kittens is male'},$$

we see that the probability that we are searching for is $P(\{(M, M)\}|E)$, i.e. the probability of having two male kittens given that one was already born.

By Bayes' theorem we have

$$P(\{(M, M)\}|E) = P(E|\{(M, M)\})\frac{P(\{(M, M)\})}{P(E)}.$$

Since the birth of males or females in cats happens with the same frequency, we have

$$P(\{(M, M)\}) = \frac{1}{4}.$$

The likeliness $P(E|\{(M, M)\})$ is the probability that at least one of the kittens is male given that both of the kittens are male, which is clearly 1.

Finally, to compute the marginal $P(E)$ we can use the law of total probability, i.e. eq. (1.12), which says that

$$P(E) = P(E|F)P(F) + P(E|F^c)P(F^c)$$

for any other event F.

In our case, if we take $F = \{(M, M)\}$ we get:

$$P(E) = P(E|\{(M, M)\})P(\{(M, M)\}) + P(E|\{(M, M)\}^c)P(\{(M, M)\}^c),$$

but we know that $P(\{(M, M)\}) = \frac{1}{4}$ and

- $P(\{(M, M)\}^c) = \frac{3}{4}$ because we have 3 equally likely options over the 4 of the sample space
- to compute $P(E|\{(M, M)\}^c)$ we note that $\{(M, M)\}^c = \{(F, F), (F, M), (M, F)\}$, so there are 2 chances over 3 that E occurs, i.e. that at least one of the kittens is male, for this reason $P(E|\{(M, M)\}^c) = \frac{2}{3}$.

Putting together all the information gathered above we get

$$P(\{(M, M)\}|E) = 1 \cdot \frac{\frac{1}{4}}{1 \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{3}{4}} = \frac{1}{3} \simeq 33\%$$

$\diamond$

Example 1.6.5 (**Murder she wrote**) A person with 0 blood type is arrested by the police as a suspect for a murder. On the murder scene, the famous writer, murder-solver... and curse-bringer... Jessica Fletcher finds two blood traces not belonging to the victim. Once analyzed, one of them is found to be of type 0 and the other of type AB.

Knowing that 60% of people have 0 blood type and only 1% of people have AB blood type, establish if the novel information gathered by Jessica Fletcher raises or lowers the probability that the person arrested by the police is the actual murderer.

The suspect is either guilty or not, and this is a binary event, hence the increase or decrease in probability can be quantified through the Bayes factor.

As always, we must identify the pertinent *events* of this problem:

- E: 'the suspect has left a trace of his blood on the murder scene'
- z: 'blood traces of type 0 are found at the murder scene'
- ab: 'blood traces of type AB are found at the murder scene'.

The Bayes factor is

$$\mathscr{B} = \frac{P(z \cap ab|E)}{P(z \cap ab|E^c)}.$$

We have $P(z|E) = 1$, because if E occurs, i.e. the suspect left traces of his blood at the murder scene, then, since the suspect has 0 blood type, z occurs.

Now imagine that E occurs and both traces of blood of type 0 *and* AB are found, i.e. $z \cap ab$ occurs. It is clear that the occurrence of E has no influence on the even ab, i.e. E and ab are independent events. Hence,

$$P(z \cap ab|E) = P(z|E) \cdot P(ab|E) = 1 \cdot 0.01 = 0.01.$$

Suppose now that E^c occurs, i.e. that the suspect did not leave any trace of his blood on the murder scene. Then, the fact of finding two types of blood traces must be ascribed to 2 other persons with blood type 0 (with probability 0.6) and AB (with probability 0.01). So

$$P(z \cap ab|E^c) = 2 \cdot 0.01 \cdot 0.6 = 0.012.$$

To conclude, the Bayes factor gives:

$$\mathscr{B} = \frac{0.01}{0.012} \simeq 0.8 < 1.$$

Since $\mathscr{B}$ is *less than 1*, the probability that the suspect is indeed guilty has lessen after the discovery of the two blood traces, one of them being the rare blood type AB. $\diamond$

1.7 Questions about chapter 1

It is important that you check your knowledge about the following items, which are the most important ones of the first chapter. It is a very good idea to test your knowledge with at least another person or, even better, in a group. Do now limit yourself to answer the following questions, but **always try to accompany them with an example**.

Typically, you will find that you forgot some concepts, this is completely normal and simply means that you have to 'rewind the tape' and go back to those particular concepts.

After a very few iterations, chapter one should have no mysteries left for you...

1. What are the sample space and an event in probability? What does it mean that an event occur?
2. What are the event operations and, most importantly, what is their meaning?
3. How do you mathematically define two incompatible events?
4. What is a partition of the sample space?
5. How can you define the probability function?
6. Knowing the probability $P(E)$ of an event, what is the probability of the complementary event E^c?
7. If two events E and F are such that $E \subseteq F$, then what is the relation between the corresponding probabilities?
8. Do you remember what is the probability that either
9. Can you quote the fundamental theorem of counting?
10. What is a permutation of n labels?
11. Do you remember how to fill the table of the 4 possible counting of ways to choose items with or without ordering and with or without replacement?
12. Given a sample space with n equally likely outcomes, how can you define the probability of and event?
13. What is the conditional probability and how is it related to the concept of updating the sample space?
14. Do you remember the Bayes theorem for two events and the name of the probabilities involved in the theorem?
15. How can you extend it to the case in which you know a partition of the sample space?
16. What does it mean, mathematically and probabilistically, that two events are statistically independent?
17. Do you recall how to extend the definition of statistical independence to more than two events?
18. What is the meaning of odds in favor of an event and how do they relate to the Bayes ratio?

Chapter 2

Random variables and distributions

One of the main protagonists of probability is the concept of random variable, which is actually ... a deterministic function!

As always, let us motivate the definition with a concrete example. Suppose we conduct an opinion pool on a group of 50 people, asking them if they agree or not upon a certain topic. If we record a '1' for *agree* and '0' for *disagree*, then each outcome of the pool will be an ordered string of size 50 of 0's and 1's, e.g.

$$(\underset{1}{0}, \underset{2}{0}, \underset{3}{1}, \underset{4}{0}, \underset{5}{1}, \underset{6}{1}, \underset{7}{1}, \underset{8}{0}, \underset{9}{1}, \ldots, \underset{50}{0}).$$

So, the sample space S is the set of all possible such strings, whose number is huge: $2^{50} \simeq 1.126 \cdot 10^{15}$.

Of course, it would be useful to reduce S to a smaller size. A reasonable idea seems to be the following: in general, what is interesting in a pool is the *number of people who agree* out of the total sample of persons contacted. This number is exhaustive because, by exclusion, we can obtain the number of people who disagree.

So, if we pick an element of S, i.e. a string s of 50 1's and 0's, and we count how many 1's have been recorder out of 50, the essence of the pool will be captured.

This amounts at defining the function

$$\begin{array}{rccl} X: & S & \longrightarrow & \mathbb{R} \\ & s & \longmapsto & X(s) := \text{number of 1's recorded in the string } s. \end{array}$$

Notice that X is a deterministic function, i.e., once we select the string s, X will give us the *exact value* of 1's in it. However, we *do not know the information inside s until we evaluate X in it*. Hence, the randomness is intrinsically incorporated in the sample space S.

The range of X, denoted with $X(S)$, is the set of all the images of X, i.e.

$$X(S) = \{y \in \mathbb{R} \ : \ \exists s \in S \text{ such that } y = X(s)\},$$

in our case it is clear that

$$X(S) = \{0, 1, 2, \ldots, 50\},$$

which is clearly much easier to handle than the original sample space S, which is composed by a huge number of strings.

The definition of random variable is the abstract generalization of what we have just discussed.

Def. 2.0.1 (Random variable) *Let S be a sample space, then:*

- *a function $X : S \to \mathbb{R}$ is called a random variable*
- *the range of X is the set of all the images of S via X, i.e. $X(S)$, and it is denoted with $\mathcal{X} = X(S)$*
- *if $\mathcal{X}$ is a finite or countable subset of $\mathbb{R}$, then X is a **discrete random variable**, otherwise it is a **continuous random variable**.*

The notation, luckily, is standard and universally used in every book on probability:

- **random variables** are always denoted with **uppercase letters**
- **the values taken by the variable** belonging to its range $\mathcal{X} \subseteq \mathbb{R}$ are always denoted by **the corresponding lowercase letters**.

2.1 The law of a discrete random variable

Given a sample space S, a probability function P related to S and a random variable $X : S \to \mathbb{R}$, we can build a function $P_X : \mathcal{X} \to \mathbb{R}$, defined on the novel sample space $\mathcal{X} = X(S)$.

In this section we concentrate on *discrete random variables* and we will extend our analysis to continuous random variables later. To better grasp the concepts that we will introduce, let us deal with a finite simple space:

$$S = \{s_1, s_2, \dots, s_n\},$$

then the random variable X defines a novel sample space, the range of X:

$$\mathcal{X} = X(S) = \{X(s_1), X(s_2), \dots, X(s_n)\} =: \{x_1, \dots, x_m\}, \quad m \leqslant n.$$

As we have seen before in the pool example, m can be significantly smaller than n, however, if X is one-to-one, then $n = m$. The function $P_X : \mathcal{X} \to \mathbb{R}$ is defined as follows:

$$\begin{array}{rccl} P_X : & \mathcal{X} & \longrightarrow & \mathbb{R} \\ & x_i & \longmapsto & P_X(x_i) := P(E_i), \quad E_i := \{s_j \in S \,:\, X(s_j) = x_i\}), \quad i = 1, \dots, m, \end{array}$$

which means that P_X computed in x_i is the original probability P of the *event* E_i given by the *set* of all outcomes $s_j \in S$ such that $X(s_j) = x_i$.

Notation: instead of writing $P_X(x_i)$, the common notation in practically every book of probability is the following

$$P(X = x_i) = P(E_i), \quad E_i = \{s_j \in S \,:\, X(s_j) = x_i\}).$$

There are *two abuses of notation* in this convention:

- P_X looses the X and it may be confused with the original probability function P, however this is avoided by the explicit appearance of X in the argument of P
- $X = x_i$ makes no sense if interpreted in a strict mathematical sense, because it is an equality between a function, X, and a real value x_i.

Def. 2.1.1 (Law of a discrete random variable) *The law of a discrete random variable X is the following set of probability values*

$$\text{law}(X) := \{P(X = x_i),\ x_i \in \mathcal{X}\}.$$

If the number n of values of x_i is not too big, is custom to visualize the law of X through a table like the following

x_i	x_1	x_2	x_3	$\dots$	x_n
$P(X = x_i)$	p_1	p_2	p_3	$\dots$	p_n

where the values p_i must of course satisfy the constraint $p_1 + p_2 + \cdots + p_n = 1$ because P is a probability function.

Example 2.1.1 Consider again Example 1.5.6 of the three coin tosses. Define the random variable X:= *number of heads obtained in the three tosses*. The following table provides the enumeration of the value of X for each point in the sample space.

s	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(s)$	3	2	2	2	1	1	1	0

The range for the random variable X is $\mathcal{X} = \{0, 1, 2, 3\}$. Assuming that all eight points in S have probability $1/8$, by a straightforward counting in the table above, we see that the law of the random variable X is given by

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

For example, $P(X = 1) = P(\{\text{HTT, THT, TTH}\}) = \frac{3}{8}$. ◇

Let us now examine the law of a random variable relative to the example of the pool.

Example 2.1.2 It can be possible to determine the law associated to a random variable X even if a complete listing of S, as in the previous Example, is not explicitly available, e.g. because too large.

This is the case of the sample space S of the 2^{50} strings of 50 0's and 1's of the pool example that started this chapter. Let X be again the random variable given by *number of 1's*, and $\mathcal{X} = \{0, 1, 2, \dots, 50\}$.

If each of the 2^{50} strings is equally likely, the probability that $X = x \in \mathcal{X}$, e.g. $X = 19$, can be obtained by counting all of the strings with x 1's in the original sample space. Since we are dealing with an unordered enumeration without replacement, we have:

$$\text{number of strings with } x \text{ 1's} = \binom{50}{x},$$

moreover, since each string is equally likely, we obtain

$$P(X = x) = \frac{\text{number of strings with } x \text{ 1's}}{\text{number of strings}} = \frac{\binom{50}{x}}{2^{50}}.$$

◇

2.2 The cumulative distribution function of a random variable

Let us now suppose that we are dealing with a *continuous* random variable X. If we try to replicate the definition of law that we have introduced in the previous section in this case, we realize quite rapidly that this simple idea cannot work.

In fact, the probability that X takes a particular real value x is, in general, 0. To understand this seemingly strange sentence, imagine that you want to compute the probability that the level of cholesterol in your blood is $x = 0.539692529821739\ldots\ mg/\ell$, i.e. $P(X = x)$.

Since the real value x has infinite digits, it is impossible to know the level of cholesterol with such a perfect precision, hence the probability that X gives rise to such a particular value is null, i.e. $P(X = x) = 0$.

This explains why, for continuous random variable, we are not interested at the expression $P(X = x)$, but at the expression $P(a \leqslant X \leqslant b)$, i.e. the probability that the random variable X takes a value that falls in the interval $[a, b]$, which is much more likely to give a meaningful result:

$$P(a \leqslant X \leqslant b) = P(\{s \in S \ : \ X(s) \in [a, b]\}).$$

We formalize this expression with the useful concept of cumulative distribution function (or *fonction de répartition* in French), for both continuous and discrete probabilities.

Def. 2.2.1 (CDF) *Given a discrete or continuous random variable X, its cumulative distribution function* (CDF) *is the function F_X defined as follows:*

$$\begin{array}{rccl} F_X: & \mathbb{R} & \longrightarrow & [0,1] \\ & x & \longmapsto & F_X(x) := P(X \leqslant x). \end{array}$$

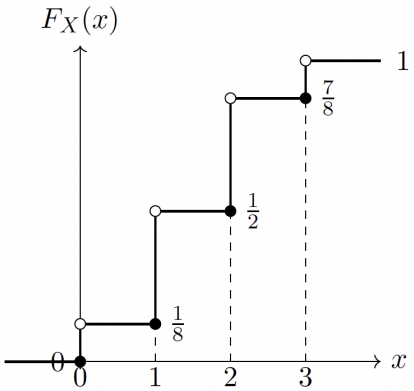
So, $F_X(x)$ gives the probability that the values expressed by X are less or equal to x.

Example 2.2.1 (CDF of tossing three coins) Consider again the experiment of tossing three fair coins, and let $X :=$ *number of heads observed*. We already know from Example 2.1.1 that

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

So, the CDF of X is a so-called *step function* given by

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 0 + \frac{1}{8} = \frac{1}{8} & \text{if } 0 \leqslant x < 1 \\ \frac{1}{8} + \frac{3}{8} = \frac{1}{2} & \text{if } 1 \leqslant x < 2 \\ \frac{1}{2} + \frac{3}{8} = \frac{7}{8} & \text{if } 2 \leqslant x < 3 \\ \frac{7}{8} + \frac{1}{8} = 1 & \text{if } 3 \leqslant x < +\infty \end{cases}$$



$\diamond$

Whether a CDF is continuous or has jumps corresponds to the associated random variable being continuous or not. In fact, the following result holds.

Theorem 2.2.1 *A random variable X is discrete if and only if F_X is a step function of x, and it is continuous if and only if F_X is a continuous function of x.*

A CDF of a random variable X can be fully characterized by the properties listed below.

Theorem 2.2.2 (Characterization of CDF) *A function $F : \mathbb{R} \to [0, 1]$ is the* CDF *of a random variable if and only if*

1. $\lim_{x \to -\infty} F(x) = 0$
2. $\lim_{x \to +\infty} F(x) = 1$
3. *F is non-decreasing: for all $x, y \in \mathbb{R}$ such that $x < y$, we have $F(x) \leqslant F(y)$.*

Let us use this result to show an example of *continuous* CDF.

Example 2.2.2 Let us check if the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = \frac{1}{1 + e^{-x}}$$

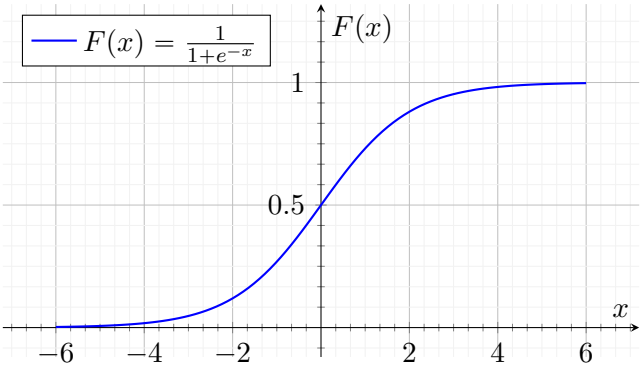
is a CDF.

- $e^{-x} \underset{x \to -\infty}{\to} +\infty$, so $\lim_{x \to -\infty} F(x) = 0$
- $e^{-x} \underset{x \to +\infty}{\to} 0$, so $\lim_{x \to +\infty} F(x) = 1$
- Recall that if a function of a real variable is differentiable and its first derivative is $\geqslant 0$, then it is non-decreasing. Let us derive F:

  $$F'(x) = \frac{e^{-x}}{(1 + e^{-x})^2} > 0,$$

  which exists and it is positive for all $x \in \mathbb{R}$, and F is strictly increasing.

So, F is an example of continuous CDF, called **logistic**, whose graph is the *sigmoid* curve depicted below.



◇

2.2.1 Identically distributed random variables

We close this section with a theorem which says that a random variable X is completely determined by its CDF F_X. To state this result in a precise way, we need to understand what it means for two random variables to be identically distributed.

Def. 2.2.2 (Identically distributed random variables) *The random variables X and Y are identically distributed if, for every subset $A \subseteq \mathbb{R}$ we have*

$$P(X \in A) = P(Y \in A).$$

It is very important to stress that *two random variables that are identically distributed are not necessarily equal*, i.e. being identically distributed does not mean $X = Y$.

Example 2.2.3 (**Identically distributed, but not equal, random variables**)
Consider again the random experiment of tossing a fair coin three times and define the random variables X and Y by:

$$X: \text{ number of heads observed}$$

$$Y: \text{ number of tails observed.}$$

From Example 2.1.1 we know that

s	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT

since s is distributed symmetrically with respect to heads and tails we have:

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

and

x	0	1	2	3
$P(Y = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Since $P(X = x) = P(Y = y)$, X and Y are identically distributed, however for all s, $X(s) \neq Y(s)$, hence $X \neq Y$! $\diamond$

Now we can state the theorem announced at the beginning of this section.

Theorem 2.2.3 *The random variables X and Y are identically distributed if and only if $F_X(x) = F_Y(x)$ for every $x \in \mathbb{R}$.*

The proof of this theorem is not simple and it is omitted.

2.3 Density and mass function of a random variable

Associated with a random variable X and its CDF F_X is another function, called either the probability mass function (PMF), or the probability density function (PDF), depending on the fact that X is discrete or continuous, respectively.

The PMF is concerned with *point probabilities* of random variables, in the following sense.

Def. 2.3.1 (PMF) *The probability mass function of a discrete random variable X is the function $f_X : \mathbb{R} \to [0, +\infty)$ given by*

$$f_X(x) := P(X = x),$$

for all $x \in \mathbb{R}$.

The passage from the discrete to the continuous case is not that simple. As we have already said, if we naively try to calculate $P(X = x)$ for a continuous random variable we end up with 0. So, instead of considering point probabilities we must deal with intervals.

Def. 2.3.2 (PDF) *If it exists a function $f_X : \mathbb{R} \to [0, +\infty)$ such that*

$$P(a \leqslant X \leqslant b) = \int_a^b f_X(x)dx, \qquad \text{for all } a, b \in \mathbb{R},\ a \leqslant b,$$

then it is called the probability density function (or simply density) of the random variable X.

The normalization of probability imposes that

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

If X has a density, then, we can understand very easily why, for every $a \in \mathbb{R}$, the probability that X takes the particular value a is 0, in fact

$$P(X = a) = P(a \leqslant X \leqslant a) = \int_a^a f_X(x)dx = 0.$$

All random variables X have CDF but they do not necessarily have densities. However, if the CDF F_X is differentiable, then X has indeed a density which can be obtained from F_X.

Theorem 2.3.1 *If the* CDF F_X *of a continuous random variable X is differentiable, then X has a density f_X that is the first derivative of F_X:*

$$f_X = F_X'.$$

Applying the fundamental theorem of calculus to the formula $f_X = F_X'$, we find the integral relationship between the PDF f_X and the CDF F_X: for all $x \in \mathbb{R}$

$$F_X(x) = \int_{-\infty}^x f_X(t)dt,$$

so

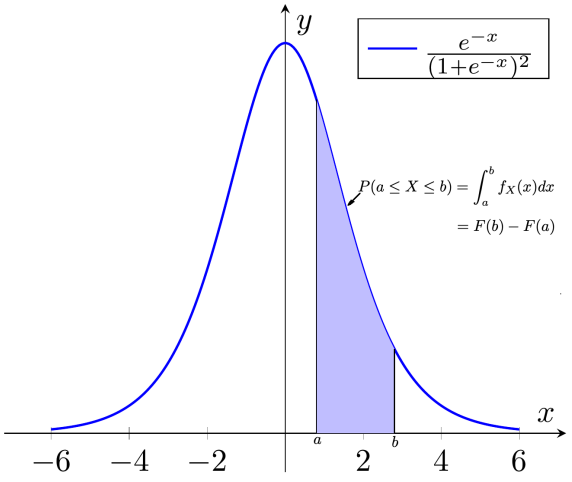
$$P(a \leqslant X \leqslant b) = F_X(b) - F_X(a).$$

Example 2.3.1 (**Logistic probabilities**) Let us consider again the logistic distribution that we have analyze in Example 2.2.2:

$$F_X(x) = \frac{1}{1+e^{-x}},$$

then

$$f_X(x) = F'(x) = \frac{e^{-x}}{(1+e^{-x})^2}.$$



It is well-known that the geometric meaning of the integral of a positive-valued function is the area underneath the graph of this function, so we have

- $F_X(b)$: area underneath the graph of f_X when $x \in (-\infty, b]$
- $F_X(a)$: area underneath the graph of f_X when $x \in (-\infty, a]$
- $F_X(b) - F_X(a)$: area underneath the graph of f_X when $x \in [a, b]$.

◇

The only two requirements for a PMF or a PDF are expressed in the following theorem.

Theorem 2.3.2 *A function f_X is a* PMF *or* PDF *of a random variable X if and only if*

1. $f_X(x) \geqslant 0$ *for all* $x \in \mathbb{R}$
2. $\sum_x f_X(x) = 1$ *for a* PMF *and* $\int_{-\infty}^{+\infty} f_X(x) = 1$ *for a* PDF.

The expression 'X has a distribution given by F_X' is abbreviated symbolically by $X \sim F_X$, where we read the symbol $\sim$ as 'is distributed as'.

We can similarly write $X \sim f_X$, or, if X and Y have the same distribution, $X \sim Y$.

2.4 Expectation value, variance and standard deviation of a random variable

In this section we introduce very important concepts related to random variables. We voluntarily postpone the examples to the following section, when we will introduce the most common densities in probability.

2.4.1 Expectation value (or mean)

The expectation value of a random value is simply its **weighted average**, with weights given by its probability density distribution.

To make this concept clearer let us consider this set of 20 of outcomes of rolling a die:

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

We can compute the average number by summing all the outcomes and dividing by 20, which gives 3.35. Or, we can calculate the same average by observing that:

- 1,2 and 6 appear 4 times
- 3 and 4 appear 3 times
- 5 appears 2 times.

So, we can write the average A as

$$A = \frac{1\cdot 4 + 2\cdot 4 + 6\cdot 4 + 3\cdot 3 + 4\cdot 3 + 5\cdot 2}{20} = \frac{(1+2+6)\cdot 4 + (3+4)\cdot 3 + 5\cdot 2}{20},$$

but we can rearrange this formula as follows

$$A = (1+2+6)\frac{4}{20} + (3+4)\frac{3}{20} + 5\frac{2}{20} = 9\cdot\frac{1}{5} + 7\cdot\frac{3}{20} + 5\cdot\frac{1}{10} = \frac{9}{5} + \frac{21}{20} + \frac{1}{2} = 3.35.$$

The numbers

- $w_1 = 1/5$
- $w_2 = 3/20$
- $w_3 = 1/10$

are the weights of the weighted sum. They have two key properties:

- they are all numbers between 0 and 1
- counted with the number of times that they repeat, they sum up to 1,

in fact

- $w_1 = 1/5$ repeats three times (for 1, 2 and 6)
- $w_2 = 3/20$ repeats two times (for 3 and 4)
- $w_3 = 1/10$ appears only once (for 5),

hence

$$3 \cdot \frac{1}{5} + 2 \cdot \frac{3}{20} + \frac{1}{10} = \frac{3}{5} + \frac{6}{20} + \frac{1}{10} = \frac{6}{10} + \frac{3}{10} + \frac{1}{10} = \frac{6+3+1}{10} = 1.$$

The PMF has exactly the same property and the also the PDF behaves like this, if we replace the discrete sum with an integral.

So, if we multiply the values that a discrete random variable can take by the corresponding PMF and we sum, then this will be the mean value of the random variable written under the form of a weighted average, with weights given by the PMF.

The same can be said for a continuous random variable, replacing the PMF with the PDF and the sum with the integral.

As always, people working in probability like having their own definition and symbol...

Def. 2.4.1 (Expectation value - or mean - of a discrete random variable) *Given the discrete random variable X, its expectation value is*

$$E[X] := \sum_x x f_X(x),$$

where f_X is the PMF *of X.*

Def. 2.4.2 (Expectation value - or mean - of a continuous random variable) *Given the continuous random variable X with* PDF *f_X, its expectation value is*

$$E[X] := \int_{-\infty}^{+\infty} x f_X(x) dx,$$

if this value if finite.

Theorem 2.4.1 (Properties of the expectation value) *Let $a, b \in \mathbb{R}$ and X, Y random variables, then*[1]

1. $E[a] = a$
2. $E[aX] = aE[X]$
3. $E[X + a] = E[X] + a$
4. $E[aX + bY] = aE[X] + bE[Y]$
5. $E[XY] = E[X]E[Y]$ *if and only if X and Y are independent random variables*[2].

The first four properties descend from the linearity of the sum or the integral and by the normalization of the density. The last one is more complicated to prove.

[1] a used as a random variable denotes $X : S \to \mathbb{R}$, $s \mapsto X(s) = a$ for all s, i.e. the a-constant random variable.

[2] knowing the value that X takes in $s \in S$ gives no information about the value that Y takes in s. The formalization of this concept needs the introduction of the joint and marginal probability distributions, which would take us too far.

2.4.2 Moments

The so-called moments of a random variable X, which sometimes are referred to its distribution F_X, are special classes of expectation values.

Def. 2.4.3 (n-th moment) *Let $n \geqslant 1$ be an integer. Then the n-th moment of X is the expectation value of the n-th power of X:*

$$\tilde{\mu}(X) := E[X^n].$$

Notice that *the 1-th moment of X is its expectation value.*

Alongside the moment, it is useful to define a closely related quantity which can be obtained by shifting it by the expectation value of X.

Def. 2.4.4 (n-th centered moment) *Let $n \geqslant 1$ be an integer. Then the n-th centered moment of X is*

$$\mu(X) := E[(X - E[X])^n].$$

Among all the possible moments of higher order, i.e. $n \geqslant 2$, the most important is the second order central moment, which bears a special name.

Def. 2.4.5 (Variance and standard deviation) *The second central moment of X is called its variance:*

$$\mathrm{Var}(X) := E[(X - E[X])^2]$$

The positive square root of $\mathrm{Var}(X)$ is called standard deviation of X, written

$$\sigma(X) := \sqrt{\mathrm{Var}(X)}.$$

The **variance** gives a **measure of the degree of spread of a distribution around its expectation value**: large values of the variance mean that X varies a lot, while, at the other extreme, if $\mathrm{Var}(X) = E[(X - E[X])^2] = 0$, then X is equal to $E[X]$ with probability 1, so there is no variation in X at all.

The qualitative interpretation of the standard deviation is the same at that of the variance, the only difference between the two is that *the unit of measurement of the standard deviation is the same as that of X*, while that of the variance is the square of the original unit of X.

It is sometimes easier to use an alternative formula for the variance, given by

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2,$$

the equivalence between the definition and this formula can be easily proven simply by developing the square $(X - E[X])^2$:

$$\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X]^2],$$

using the linearity of the expectation value, i.e. property 4. of theorem 2.4.1, we have

$$\mathrm{Var}(X) = E[X^2] - E[2XE[X]] + E[E[X]^2],$$

but $E[X]$ and $E[X]^2$ are constants, so, by properties 1. and 2 of theorem 2.4.1 we can simplify the previous formula as

$$\mathrm{Var}(X) = E[X^2] - 2E[X]E[X] + E[X]^2 = E[X^2] - 2E[X]^2 + E[X]^2 = E[X^2] - E[X]^2.$$

The choice to use one formula or the other depends on which one provides an easier integral to calculate in the explicit formulae that follow:

$$\begin{aligned}\mathrm{Var}(X) &= E[(X - E[X])^2] = \int_{-\infty}^{+\infty} (x - E[X])^2 f_X(x)dx \\ &= E[X^2] - (E[X])^2 = \int_{-\infty}^{+\infty} x^2 f_X(x)dx - \left(\int_{-\infty}^{+\infty} x f_X(x)dx\right)^2.\end{aligned}$$

Let us now see how the variance behaves when we perform an affine transformation of a random variable.

Theorem 2.4.2 *If X is a random variable with finite variance, then for any constants a and b it holds that*

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X).$$

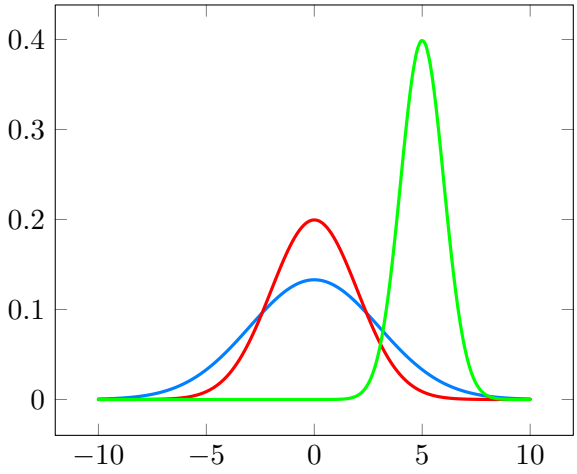
Proof. By definition of variance we have:

$$\begin{aligned}\mathrm{Var}(aX + b) &= E[(aX + b - E[aX + b])^2] = E[(aX + \not{b} - aE[X] - \not{b})^2] \\ &= E[(aX - aE[X])^2] = E[a^2(X - E[X])^2] = a^2E[(X - E[X])^2] \\ &= a^2\mathrm{Var}(X).\end{aligned}$$

□

In particular, note that $\mathrm{Var}(\lambda) = 0$ for all constant random variable, $\lambda \in \mathbb{R}$.

Example 2.4.1 Let us interpret the PDFs depicted in the following figure.



- Blue PDF: symmetric and picked in 0, so its expectation value is 0, but due to the large spread, the variance associated to this PDF is the largest among the three depicted.
- Red PDF: symmetric and picked in 0, the expectation value is again 0 but more probable than for the blue PDF because it is less spread around 0, so its variance is smaller.
- Green PDF: also symmetric, but picked in 5, so its expectation value is 5 and this with a quite high probability, because the spread around that value is the smallest among the three curves, so its variance is also the smallest one.

◇

2.5 Common density distributions

In this section we catalog many of the more common statistical distributions for discrete and continuous random variables. For each distribution we will give its mean and variance and many other useful or descriptive measures that may aid understanding. We will also indicate some typical applications of these distributions.

Note that this section is by no means comprehensive in its coverage of statistical distributions.

2.5.1 Discrete density distributions

We recall that a random variable X has a discrete distribution if the range of X, the sample space, is a finite or countable set.

Discrete uniform distribution

A random variable X has a discrete uniform distribution if its PMF is

$$P(X = x) = \frac{1}{N}, \qquad x = 1, \ldots, N.$$

This distribution puts *equal mass* on each of the outcomes $1, 2, \ldots, N$.

To calculate the expectation value and variance of X, we must use the identities

$$\sum_{j=1}^{N} j = \frac{N(N+1)}{2}, \quad \sum_{j=1}^{N} j^2 = \frac{N(N+1)(2N+1)}{6}.$$

- $E[X]$:

$$E[X] = \sum_{x=1}^{N} xP(X = x) = \frac{1}{N}\sum_{x=1}^{N} x = \frac{N(N+1)}{2N} = \frac{N+1}{2},$$

so, the **expectation value of the uniform distribution** is

$$E[X] = \frac{N+1}{2}.$$

- $\mathrm{Var}(X)$:

$$E[X^2] = \sum_{x=1}^{N} x^2 P(X = x) = \frac{1}{N}\sum_{x=1}^{N} x^2 = \frac{N(N+1)(2N+1)}{6N} = \frac{(N+1)(2N+1)}{6},$$

hence

$$E[X^2] - (E[X])^2 = \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} = \frac{N^2-1}{12},$$

so, the **variance of the uniform distribution** is

$$\mathrm{Var}(X) = \frac{N^2-1}{12}.$$

Bernoulli distribution

This distribution is based on the idea of a **Bernoulli trial**. A Bernoulli trial (named for the Swiss mathematician Jacob Bernoulli (1655-1705), one of the founding fathers of probability theory) is an experiment with two, and only two, possible outcomes.

A random variable X has a p-Bernoulli distribution if

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}.$$

- The value $X = 1$ means *success* and p is referred to as the *success probability*.
- The value $X = 0$ means *failure*.

The **expectation value** and **variance** of a **p-Bernoulli random variable** are

$$\begin{cases} E[X] = p \\ \mathrm{Var}(X) = p(1-p) \end{cases}.$$

To verify the first formula we recall that, by definition, to compute the expectation value we have to sum the values taken by X multiplied by their distribution. But since X can only take values 1 and 0 with probabilities p and $1-p$, respectively, we have:

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

Repeating the same reasoning for the variance and using the fact that $E[X] = p$ we find:

$$\begin{aligned} \mathrm{Var}(X) &= E[(X - E[X])^2] = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) \\ &= p(1-p)^2 + p^2(1-p) = p(1-p)(1-p+p) = \\ &= p(1-p). \end{aligned}$$

Many experiments can be modeled as a sequence of Bernoulli trials: the repeated tossing of a coin, gambling games (e.g. roulette), election polls and several others.

Binomial distribution

The binomial distribution is closely related to the Bernoulli one. In a sequence of n identical, independent Bernoulli trials, each with success probability p, define the random variables $X_1, \dots, X_n$ by

$$X_j := \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}.$$

Then, the random variable

$$X := \sum_{j=1}^{n} X_j$$

can be proven to have the distribution given by the following PMF

$$P(X = x|n,p) := \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 1, 2, \dots, n,$$

which is called a (n,p)-**binomial distribution**.

The **expectation value** and **variance** of a **binomial** (n,p) **random variable** are

$$\begin{cases} E[X] = np \\ \mathrm{Var}(X) = np(1-p) \end{cases}.$$

We prove just the first formula, the second proof being analogous but longer. First of all we must note that

$$x\binom{n}{x} = \cancel{x}\frac{n!}{\cancel{x}(x-1)!(n-x)!} = \frac{n(n-1)!}{(x-1)!(n-1+1-x)!} = n\frac{(n-1)!}{(x-1)!((n-1)-(x-1))!},$$

i.e.

$$x\binom{n}{x} = n\binom{n-1}{x-1}. \tag{2.1}$$

To compute the expectation value we can start summing from $x = 1$ because the term for $x = 0$ is null:

$$E[X] = \sum_{x=1}^{n} xP(X = x|n,p) = \sum_{x=1}^{n} x\binom{n}{x} p^x p^{n-x} = \sum_{x=1}^{n} n\binom{n-1}{x-1} p^x p^{n-x}.$$

A property of summations is that if we subtract any integer from the summation index, we have to add the same integer inside the sum. Let this integer be 1, then

$$E[X] = n\sum_{x=1-1}^{n-1} \binom{n-1}{x-1+1} p^{x+1} p^{n-(x+1)} = n\sum_{x=0}^{n-1} \binom{n-1}{x} p p^x p^{(n-1)-x},$$

p can be taken out of the sum to obtain

$$E[X] = np\sum_{x=0}^{n-1} \binom{n-1}{x} p^x p^{(n-1)-x} = np \cdot 1,$$

where the last step follows from the fact that

$$\sum_{x=0}^{n-1} \binom{n-1}{x} p^x p^{(n-1)-x}$$

is the sum over all possible values of a binomial $(n-1,p)$ probability density mass!

Example 2.5.1 (**Dice probabilities**) We are interested in the probability of obtaining at least one 6 in *four rolls* of a fair die.

This experiment can be modeled as a sequence of four Bernoulli trials with success probability

$$p = \frac{1}{6} = P(\text{die shows } 6).$$

Define the random variable X by

$$X = \textit{'total number of 6 in four rolls'},$$

then $X \sim \text{binomial}(4, 1/6)$.

The probability of obtaining at least one 6 is the probability that $X > 0$, but since $X \geqslant 0$, we can more easily compute this probability by noticing that $P(X > 0) = 1 - P(X = 0)$, so

$$P(\text{at least one } 6) = P(X > 0) = 1 - P(X = 0) = 1 - \binom{4}{0}\left(\frac{1}{6}\right)^0\left(1 - \frac{1}{6}\right)^{4-0} = 1 - \left(\frac{5}{6}\right)^4 \simeq 0.52.$$

Now we consider another game: throw a pair of dice 24 *times* and ask for the probability of at least one *double* 6. This, again, can be modeled by the binomial distribution with success probability p, where

$$p = P(\text{at least a double } 6) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

So, if $Y = \textit{'number of double 6 in 24 rolls'}$, then $Y \sim \text{binomial}(24, 1/36)$ and

$$P(\text{at least one double } 6) = P(Y > 0) = 1 - P(Y = 0) = 1 - \binom{24}{0}\left(\frac{1}{36}\right)^0\left(\frac{35}{36}\right)^{24} \simeq 0.49.$$

This is the calculation originally done in the eighteenth century by Pascal at the request of the gambler de Méré, who thought both events had the same probability. He began to believe he was wrong when he started losing money on the second bet... ◇

Poisson distribution

The Poisson distribution is a widely applied discrete distribution and can serve as a model for a number of different types of experiments. For example, it can be used to model a phenomenon in which, for small time intervals, **the probability of an occurrence is proportional to the length of waiting time**.

This is the case for the probability that a bus arrives at a stop, a customer enters in a shop, a photon is captured by a telescope, a radioactive sample decays, and many others.

Another area of application is in spatial distributions, where, for example, the Poisson model may be used to for the distribution of fish (*poisson* in French...) in a lake or the number of stars in a unit of space.

This distribution bears its name from the French mathematician Siméon Denis **Poisson** (1781-1840) who published it in 1837, even though the other French mathematician Abraham **de Moivre** (1667-1754) already found it in 1711, 126 years before!

The Poisson distribution has one parameter $\lambda \in \mathbb{R}$, $\lambda > 0$, sometimes called *intensity*.

A **random variable** X, taking values in the non-negative integers, has **a Poisson-λ distribution** if its PMF is given by

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

Quite surprisingly, **the parameter λ gives both the expectation value and the variance of X!** In fact

$$E[X] = \mathrm{Var}(X) = \lambda.$$

Let us check it for the expectation value, the proof for the variance is similar.

$$E[X] = \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} \cancel{x}\frac{e^{-\lambda}\lambda\lambda^{x-1}}{\cancel{x}(x-1)!} = \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!},$$

if now we add 1 to x inside the sum, we must subtract 1 from the summation index, obtaining

$$E[X] = \lambda e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!},$$

but the infinite sum (series) that appears in the previous expression is known to converge to the exponential of λ, so

$$E[X] = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

Example 2.5.2 (**Waiting time**) As an example of a waiting-for-occurrence application, consider a telephone operator who, on the average, handles five calls every 3 minutes. What is the probability that there will be no calls in the next minute? What about at least two calls?

If we define the random variable $X =$ '*number of calls in a minute*', then

$$P(\text{no calls next minute}) = P(X = 0) = \frac{e^{-5/3}\left(\frac{5}{3}\right)^0}{0!} = e^{-5/3} \simeq 0.19 = 19\%,$$

$$P(\geqslant 2 \text{ calls next minute}) = P(X \geqslant 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.19 - \frac{e^{-5/3}\left(\frac{5}{3}\right)^1}{1!} \simeq 0.5.$$

$\diamond$

Calculation of Poisson probabilities can be done rapidly by using the following recursion relation.

Theorem 2.5.1 *If the random variable X follows a λ-Poisson distribution, then*

$$P(X = x|\lambda) = \frac{\lambda}{x} P(X = x - 1|\lambda), \quad x = 1, 2, \ldots$$

Before proving this result, let us confirm it using the last example, in which $E[X] = 5/3 = \lambda$. The recursive formula says

$$P(X = 1|5/3) = \frac{\lambda}{1} P(X = 1 - 1|5/3) = \frac{5}{3} P(X = 0|5/3) = \frac{5}{3} e^{-5/3},$$

exactly as previously computed.

Proof. Let us write the λ-Poisson probability for $X = x$:

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-\lambda}\lambda\lambda^{x-1}}{x(x-1)!} = \frac{\lambda}{x}\frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \frac{\lambda}{x} P(X = x - 1|\lambda).$$

□

Geometric distribution

The binomial distribution counts the episodes of success in a fixed number of Bernoulli trials. Suppose that, instead, we count the number of Bernoulli trials required until we get a success. This sort of 'inverse' formulation leads to the geometric distribution.

In a sequence of independent Bernoulli-p trials, consider the following random variable $X =$ '*the trial at which the first success occurs*'.

The event $P(X = x)$ can occur only if there are no successes in the first $x - 1$ trials, and a success on the x-th trial. The probability of 0 successes in $x - 1$ trials is the binomial probability

$$\binom{x-1}{0} p^0 (1-p)^{x-1} = 1 \cdot 1 \cdot (1-p)^{x-1},$$

and with probability p there is a success on the x-th trial. Multiplying these probabilities gives **the PMF of a geometric random variable**:

$$P(X = x|p) = p(1-p)^{x-1}.$$

Since X is the trial at which the first success occurs, in probability one often says that X represents the fact that **we are waiting for a success**.

The reason for the name 'geometric' follows from the fact that, in order to prove that the sum of $P(X = x)$ as x varies from 1 to ∞ gives 1 relies on the so-called geometric series:

$$\sum_{x=0}^{\infty} q^x = \frac{1}{1-q}, \qquad \text{for all } |q| < 1,$$

more generally,

$$\sum_{x=n}^{\infty} q^x = \sum_{x=n-n}^{\infty} q^{x+n} = \sum_{x=0}^{\infty} q^n q^x = q^n \sum_{x=0}^{\infty} q^x = \frac{q^n}{1-q} \iff \sum_{x=n}^{\infty} q^x = \frac{q^n}{1-q}. \tag{2.2}$$

Since $0 < 1 - p < 1$, we have

$$\sum_{x=1}^{\infty} p(1-p)^{x-1} = p \sum_{x=1-1}^{\infty} (1-p)^{x-1+1} = \sum_{x=0}^{\infty} (1-p)^x = \frac{p}{1-(1-p)} = \frac{p}{p} = 1.$$

It can be proven that

$$\begin{cases} E[X] = \frac{1-p}{p} \\ \mathrm{Var}(X) = \frac{1-p}{p^2} \end{cases}. \tag{2.3}$$

Example 2.5.3 (**Lifetime of devices**)

The geometric distribution is sometimes used to model the lifetimes, or time until failure, of devices. For example, if the probability that a light bulb will fail on any given day is 0.001, then the probability that it will last at least 30 days is

$$P(X > 30) = \sum_{x=31}^{\infty} 0.001(1-0.001)^{x-1} = 0.001 \sum_{x=31-1}^{\infty} (1-0.001)^{x-1+1} = 0.001 \sum_{x=30}^{\infty} (1-0.001)^x$$

$$\underset{(2.2)}{=} 0.001 \frac{(1-0.001)^{30}}{1-(1-0.001)} = \cancel{0.001} \frac{(0.999)^{30}}{\cancel{0.001}} = (0.999)^{30} = 0.97 = 97\%$$

$\diamond$

2.5.2 Continuous density distributions

Now we will discuss some of the more common families of continuous distributions, those with well-known names, but keep in mind that the distributions mentioned here by no means constitute all of the distributions used in statistics.

Uniform continuous distribution

The continuous uniform distribution is defined by spreading mass uniformly over an interval $[a, b]$. Its PDF is given by

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

The normalization is immediate to prove:

$$\int_{-\infty}^{+\infty} f(x)dx = \int_a^b \frac{1}{b-a}dx = \frac{1}{b-a}(b-a) = 1.$$

It can be checked that the CDF of the uniform distribution is:

$$F_X(x) = \int_{-\infty}^{x} f(x)dx = \begin{cases} 0 & \text{if } x \leqslant a \\ \frac{x-a}{b-a} & \text{if } x \in [a, b] \\ 1 & \text{if } x \geqslant b \end{cases}.$$

So, between a and b, F_X behaves linearly with a slope $1/(b-a)$. In Figure 2.1 we present an explicit example of uniform distribution and its cumulative function.

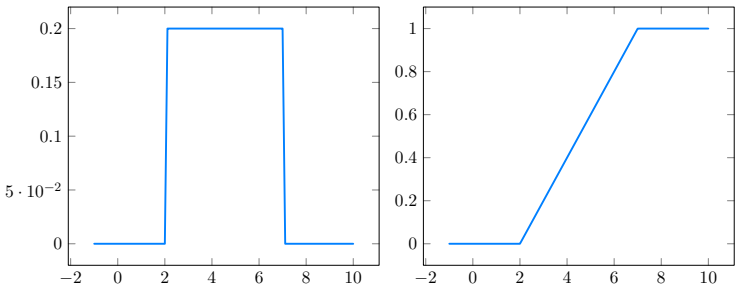


Figure 2.1: *Left*: PDF of the uniform distribution in the interval $[2, 7]$. *Right*: its CDF.

Regarding the expectation value and the variance we have:

$$\begin{cases} E[X] = \dfrac{b+a}{2} \\ \text{Var}(X) = \dfrac{(b-a)^2}{12} \end{cases}.$$

Let us verify this claim.

$$E[X] = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2},$$

$$\mathrm{Var}(X) = \int_a^b \frac{\left(x-\frac{b+a}{2}\right)^2}{b-a} dx = \frac{1}{b-a}\int_{a-\frac{b+a}{2}}^{b-\frac{b+a}{2}} x^2 dx = \frac{1}{b-a}\int_{\frac{a-b}{2}}^{\frac{b-a}{2}} x^2 dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_{\frac{a-b}{2}}^{\frac{b-a}{2}},$$

the evaluation in the upper value gives

$$\frac{1}{b-a}\cdot\frac{1}{3}\cdot\frac{(b-a)^3}{8} = \frac{(b-a)^2}{24},$$

the evaluation in the lower value gives

$$\frac{1}{b-a}\cdot\frac{1}{3}\cdot\frac{(a-b)^3}{8} = \frac{1}{b-a}\frac{-(b-a)^3}{24} = -\frac{(b-a)^2}{24},$$

subtracting what we obtained we get

$$\mathrm{Var}(X) = \frac{(b-a)^2}{24} + \frac{(b-a)^2}{24} = \frac{(b-a)^2}{12}.$$

Exponential distribution

It can be considered as the continuous version of the geometric distribution and, as such, it is frequently used to model the waiting time before an occurrence.

The continuous random variable X follows an exponential distribution with parameter $a > 0$ if its PDF is

$$f_X(x) = \begin{cases} ae^{-ax} & \text{if } x \geqslant 0, \\ 0 & \text{otherwise.} \end{cases}$$

Its CDF is

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} 1 - e^{-ax} & \text{if } x \geqslant 0, \\ 0 & \text{otherwise.} \end{cases}$$

If the distribution is interpreted as waiting time, then a has the meaning of *inverse of the average waiting time*.

In Figure 2.2 we present an explicit example of exponential distribution and its cumulative function.

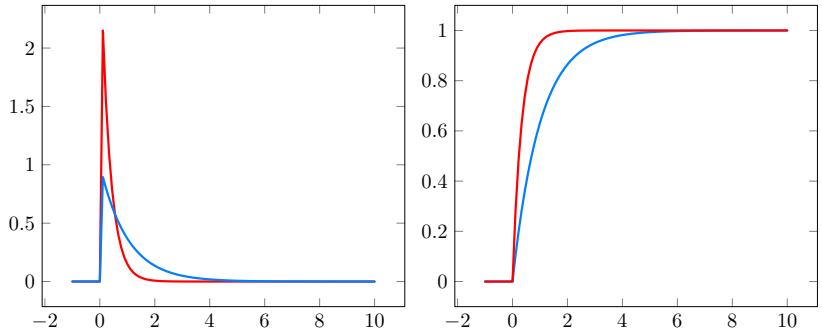


Figure 2.2: *Left*: PDF of the exponential distribution for $a = 1$ (cyan) and $a = 3$ (red). *Right*: its CDF.

The expectation value and variance of the exponential distribution are

$$\begin{cases} E[X] = \dfrac{1}{a} \\ \mathrm{Var}(X) = \dfrac{1}{a^2} \end{cases} .$$

Normal (or Gaussian) distribution

The normal distribution (sometimes called Gaussian) plays a central role in a large body of probability and statistics. There are two main reasons for this:

- first, the normal distribution is very tractable analytically (although it may not seem so at first glance!) and has very nice symmetry properties, starting from the familiar bell shape of its graph
- second, the *Central Limit Theorem* (that will be quoted later on), says that, under mild conditions, the normal distribution can be used to approximate a large variety of distributions.

The normal distribution has two parameters, usually denoted by μ and σ^2, which turn out to be its expectation value, the *mean*, and its *variance*, respectively.

The PDF of the normal distribution with mean μ and variance σ^2 (usually denoted by $\mathcal{N}(\mu, \sigma^2)$ is given by

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

and

$$\begin{cases} E[X] = \mu \\ \mathrm{Var}(X) = \sigma^2 \end{cases}.$$

The bell-shaped graph of the **standard normal PDF** ($\mu = 0, \sigma = 1$) is shown in Figure 2.3.

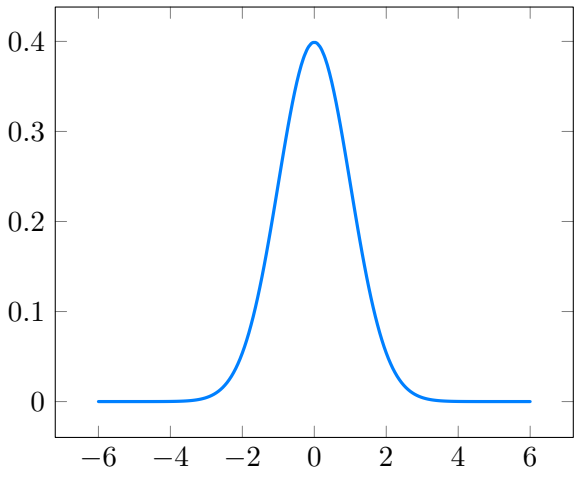


Figure 2.3: The graph of the standard normal distribution, with $\mu = 0$ and $\sigma = 1$.

The normal distribution is also special because of the fact that its two parameters, the mean μ and the variance σ^2, provide complete information about the exact shape and location of the distribution. Straightforward calculations show that the normal PDF has its *maximum* at $x = \mu$ and *inflection points* (where the curve changes from concave to convex) at $\mu \pm \sigma$.

The CDF of the normal distribution $F_X(x) = \int_{-\infty}^{x} f_X(x)dx$ cannot be written as a function, but every computer language has its values stored and available with a simple command, e.g. 'normcdf' (for Matlab) or 'pnorm' (for R).

Figure 2.4 shows how the normal distribution changes with the parameters.

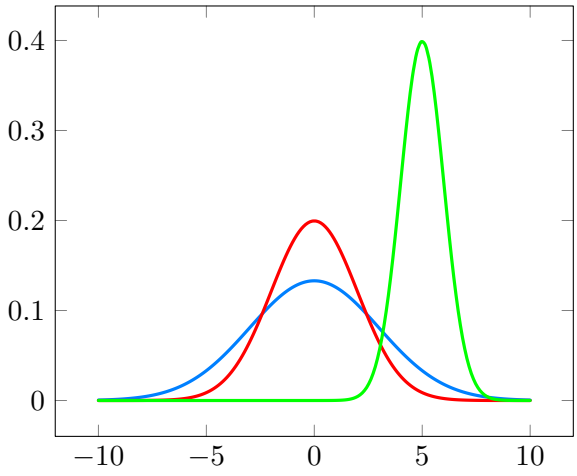
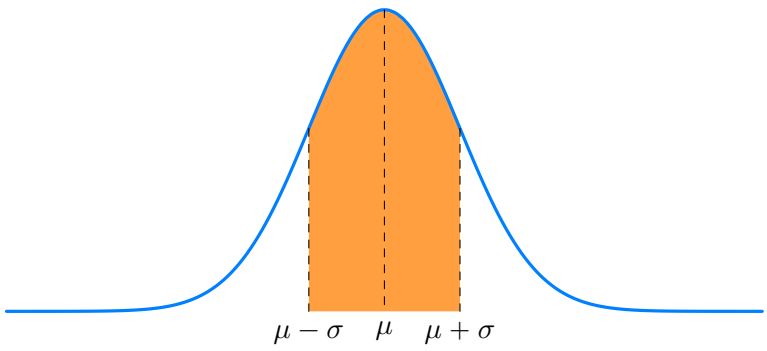


Figure 2.4: Graphs of the normal distributions with $(\mu = 0, \sigma = 3)$ (blue), $(\mu = 0, \sigma = 2)$ (red), $(\mu = 5, \sigma = 1)$ (green).

The probability content, or mass concentration, within 1, 2, or 3 standard deviations of the mean is

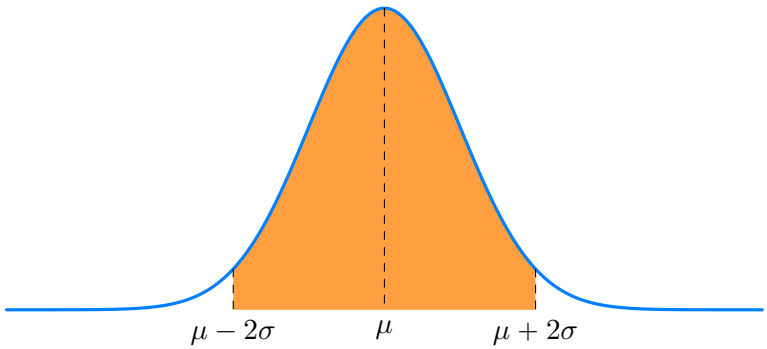
- 1 standard deviation:

$$P(\mu - \sigma \leqslant X \leqslant \mu + \sigma) \simeq 0.68 = 68\%$$



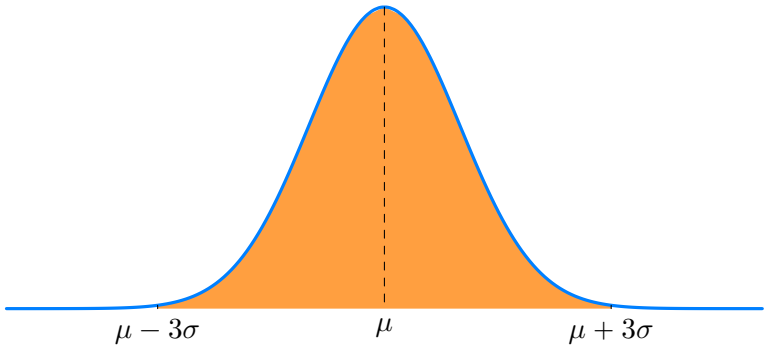
- 2 standard deviations:

$$P(\mu - 2\sigma \leqslant X \leqslant \mu + 2\sigma) \simeq 0.95. = 95\%$$



- 3 standard deviations:

$$P(\mu - 3\sigma \leqslant X \leqslant \mu + 3\sigma) \simeq 0.997 = 99.7\%$$



In the great majority of situations, a 2σ or 3σ result, i.e. the probability content of 95% or 99.7% is enough to be sure of the occurrence of a phenomenon. However, in very precise experiments, a 5σ result is considered the gold standard for significance, corresponding to about only a one-in-a-million chance that the findings are a result of random fluctuations.

An example is given by the discovery of a new particle in the LHC (large hadron collider), the accelerator at CERN in Geneva, which is the largest and most complicated machine ever built by humans.

The standardization of a normal random variable

We call a normal distribution $\mathcal{N}(\mu, \sigma^2)$ **standard** when $\mu = 0$ and $\sigma = 1$.

It is possible to transform a generic normal random variable X into a standard one through a very simple manipulation. Precisely,

$$\text{if } X \sim \mathcal{N}(\mu, \sigma^2), \text{ then } Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

This process is called *standardization* and it is useful because many statistical methods and tables are based on the standard normal distribution.

Z is often called the **standardized** random variable or the **centered and reduced** one, where 'centered' is referred to the fact that its mean μ has been set to 0 and 'reduced' to the fact that the variance has been normalized to 1.

Thanks to the symmetries of the normal distribution it's possible to show that

$$P(Z \leqslant -a) = 1 - P(Z \leqslant a), \quad \text{for all } a \in [0, 1].$$

We use this property of the standardized normal random variable in the following example.

Example 2.5.4 (**Newborn's weight**)

Let X be the random variable representing the weight of a newborn. We make the hypothesis that X follows a normal law of mean $\mu = 3.5$kg and standard deviation $\sigma = 500$g. What is the probability that a newborn has a weight less than 3.1kg?

To answer the question, let us first put the problem in standard form with the transformation just seen:

$$Z = \frac{X - \mu}{\sigma},$$

then

$$\begin{aligned} P(X \leqslant 3100) &= P\left(\frac{X - \mu}{\sigma} \leqslant \frac{3100 - \mu}{\sigma}\right) \\ &= P\left(Z \leqslant \frac{3100 - 3500}{500}\right) \\ &= P(Z \leqslant -0.8) = 1 - P(Z \leqslant 0.8) \\ &\simeq 1 - 0.79 = 0.21 = 21\% \end{aligned}$$

$\diamond$

2.6 Convergence of sequences of random variables

In the same way as we can be interested in the behavior, in particular the convergence toward a certain limit, of a sequence $(x_n)_{n\geqslant 1}$ of real numbers, we can also be interested in the behavior of a sequence of random variables $(X_n)_{n\geqslant 1}$, asking ourselves:

- What can we expect when we repeat again and again... and again a random experiment?
- Do the probability law followed by the variable for a relatively small number of experiments is the same as that for a very very big number n of experiments?
- Or again: Is there a privileged law which can approximate (in a certain sense) the other laws when $n \to +\infty$?

These are all interesting and profound questions that mathematicians tried to answer during many years and several results are at disposal. Here we will concentrate only on a result that does not need sophisticated mathematics, but only a notion of convergence that is appropriate for the random character of the variables we are interested in.

Def. 2.6.1 (Almost sure convergence) *A sequence of random variables $(X_n)_{n\geqslant 1}$ is said* ***to converge almost surely*** *to the random variable X if*

$$P\left(X_n \xrightarrow[n\to\infty]{} X\right) = 1$$

and we denote this convergence as

$$X_n \xrightarrow[n\to\infty]{a.s.} X.$$

The condition $P\left(X_n \xrightarrow[n\to\infty]{} X\right) = 1$ means that, when the number n of random experiments tends to infinity, then the probability that X_n behaves as X is 1, i.e. certain. Since probability is the most interesting feature that we are interested in, this definition makes total sense!

In the following subsections we are going to state and discuss the two most important results regarding the almost sure convergence of random variables.

2.6.1 The strong law of large numbers

Given a sequence $(X_n)_{n\geqslant 1}$ of random variables, the strong law of large numbers deals with the concept of[3] **sample mean**, which is nothing but the arithmetic average of the first n random variables of the sequence, i.e.

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

Theorem 2.6.1 (Strong law of large numbers (Kolmogorov 1929)) *Let $(X_n)_{n\geqslant 1}$ be a sequence of*

- *independent and identically distributed (i.i.d.) random variables*
- *with the same finite expectation value, i.e. $E[X_n] = \mu$, $|\mu| < +\infty$, for all $n \geqslant 1$.*

Then, the sequence of sample means $(\overline{X}_n)_{n\geqslant 1}$ converges almost surely to μ:

$$\overline{X}_n \xrightarrow[n\to\infty]{a.s.} \mu \quad \Longleftrightarrow \quad P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

Example 2.6.1 The strong law of large numbers can be used to verify, for example, if a die is fair or not. A single roll of a fair, six-sided die produces one of the numbers $1, 2, 3, 4, 5$, or 6, each with equal probability $p = 1/6$. Therefore, the expected value of the average of the rolls is:

$$\frac{1+2+3+4+5+6}{6} = 3.5.$$

The rollings of a six-sided fair die can be considered as a sequence of i.i.d. random variables with equal finite expectation value. So, by the strong law of large numbers, if a *fair* die is rolled a very large number of times, the sample mean, i.e. the average of values obtained, must approach 3.5 with the increasing precision as more dice are rolled. Figure 2.5 shows the corresponding behavior.

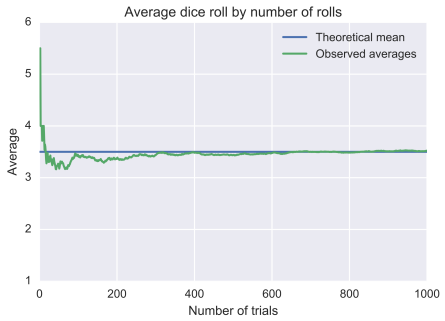


Figure 2.5: Although each run of dice would show a distinctive shape over a small number of throws (at the left), over a large number of rolls (at the right) the shapes would be extremely similar. Image: Creative Commons.

◇

[3] sometimes called **empirical mean**.

2.6.2 The central limit theorem

The strong law of large numbers says that the sample mean tends to the common expectation value of each random variable when the number of trials tends to infinity. However, it does not say anything about what kind of law is followed by the limit of the sample mean.

The central limit theorem provides the missing information.

Theorem 2.6.2 (Central limit theorem) *Let $(X_n)_{n\geqslant 1}$ be a sequence of independent and identically distributed (i.i.d.) random variables.*

If all the random variables X_n have the same expectation value $\mu = E(X_n)$ and finite variance σ, then the standardized sample mean

$$\frac{\overline{X}_n - \mu}{\sigma}$$

converges to a random variable whose distribution is the standard normal one: $\mathcal{N}(0,1)$.

In practice, the central limit theorem means that every computation of probability performed over the standardized sample mean $\overline{X}_n$ of i.i.d. random variables with a number of samples 'sufficiently large' can be approximated by the same computation performed on the standard normal law, which, in general, is much easier.

Although the central limit theorem gives us a useful general approximation, in general we have no automatic way of knowing how good the approximation is. In fact, the goodness of the approximation is a function of the original distribution, and so it must be checked case by case.

The **normal law** has a **unique feature**: it is the only continuous law with finite variance stable with respect to its mean and the central limit theorem says that the distributions that do not have this property tend to converge to the normal one.

A little bit of history: the first proof of this theorem in the case of a Bernoulli distribution of parameter $p = 1/2$ was given by De Moivre in 1738, a generalization to any parameter was proven in 1809 by Laplace. The proof of the modern version quoted above was an effort of multiple mathematicians (**Bernstein**, **Lindeberg**, **Lévy**, **Feller**, **Kolmogorov**, and others) over the period from 1920 to 1937.

The reason why the theorem bear such a peculiar denomination is due to the fact that, in 1920, the Hungarian mathematician George **Pólya** published a paper referred to the theorem as 'central' in probability theory. However, the French school of probability interprets the word 'central' in the sense that it describes the behavior of the center of the Gaussian distribution as opposed to its tails.

Example 2.6.2 (**Galton board**)

An example of sequence of random variables that fulfill the hypotheses of the central limit theorem is given by binomial variables. There is probably no better way to provide a visualization of the central limit theorem than the Galton board (also called 'bean machine'), introduced by the British mathematician Francis **Galton** (1822-1911).

The Galton board is shown in Figure 2.6.

Figure 2.6: The Galton board.

It consists of a vertical board with intermingled rows of pins. Balls are dropped from the top and **they bounce either left or right as they hit the pins**, for this reason they can be consider binomial random variables. Eventually **they are collected into bins at the bottom**. The heights of balls columns accumulated in the bins inevitably approximate the bell curve of a standard normal distribution.

Overlaying the Pascal-Tartaglia triangle onto the pins shows the number of different paths that can be taken to get to each bin, as shown in Figure 2.7.

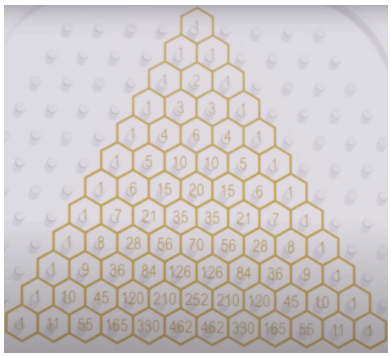


Figure 2.7: The Pascal-Tartaglia triangle superposed to the pins of the Galton board.

Consider for example the third row: there are two paths that can bring the ball to the center hexagon, and only one that can lead to the left and right extremities of the row. As we advance in the rows, the number of paths that can lead to the central positions increases dramatically with respect to those leading to the extremities. ◇

2.7 Questions about chapter 2

1. Define a random variable. When is it discrete or continuous?
2. What is the *law* of a discrete random variable?
3. Write the definition of cumulative distribution function (CDF) of a random variable. What are its 3 main properties?
4. What does it mean that two random variables are *identically distributed*?
5. Define the probability mass function (PMF) of a discrete random variable and the probability density function (PDF) of a continuous random variable.
6. If the CDF of a continuous random variable X is differentiable, what is its relation with the PDF of X?
7. Define the expectation value (or mean) for a discrete and a continuous random variable.
8. When does the following property hold true: $E[XY] = E[X]E[Y]$?
9. Define the variance and the standard deviation of a random variable. What is the meaning of the variance?
10. Can you write down the PDF of the normal distribution with mean μ and variance σ^2? When does the normal distribution is called *standard*?
11. Do you remember the percentage of probability to find a normally distributed random variable X between the mean $\mu \pm \sigma$? And $\mu \pm 2\sigma$? And $\mu \pm 3\sigma$?
12. Do you recall what it means that a sequence of random variables $(X_n)_{n\geqslant 1}$ *converges almost surely* to a random variable X?
13. What it the *sample mean* of a sequence of random variables?
14. Can you quote the strong law of large numbers? Can you explain what it means through a simple example?
15. What does it mean to standardize the sample mean?
16. Can you quote the central limit theorem and discuss its meaning?