Probability Theory Basic Concepts

—Joaquín Gómez—

—February 18, 2025—

Definition of sample space and probability law

To start with the basic concepts, we would have to define two fundamentals of probability theory in general, we must answer two questions

- When doing an experiment, or studying some event, what is the set of all outcomes that can come from that experiment/event?
- If we want to measure how likely an event is to happen, what is the **law** that assigns that "likeliness" to that event?

The first question asks for some definition of a "set of all possibilities", this is called the **sample space** and is denoted in many textbooks as Ω (big Omega). The sample space is the set of *all possible outcomes*. For example, if measuring how likely is a certain card C to be drawn from a deck of 50 Spanish cards, we would say that the sample space is

$$\Omega = \{\text{"1 of gold"}, \text{"2 of gold"}, ...\}$$

and so on. This means Ω consists of all the cards in the deck, it is the sample space of all possible cards to draw.

The second question mentions a *law* that must be defined for a particular sample space, we call this law the **probability law** and it assigns a subset $\omega \subseteq \Omega$ to a positive real number between $[0, 1]$. So, summarizing:

$$\text{The sample space } \Omega \text{ is the set of all possible outcomes of an experiment}$$
$$\text{A probability law } P(X) : \omega \to S, \text{ where } S = \{s \in \mathbb{R} : 0 \leq s \leq 1\} \text{ and } \omega$$
$$\text{is a set contained in the sample space.}$$

Axioms

There are several axioms for sample spaces and probability laws. These are:

a. $0 \leq P(\omega) \leq 1, \omega \subseteq \Omega$. The probability of any event in the sample space is always positive and less than 1.

b. $P(\Omega) = 1$, this can be interpreted as: there's a 100% probability of any event in the sample space happening.

c. $P(\emptyset) = 0$, the empty set $\emptyset$ is called an "impossible event".

(a) is very intuitive and it has been mentioned before. To prove (c) we use (b) and note

$$P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset) = 1 + P(\emptyset)$$

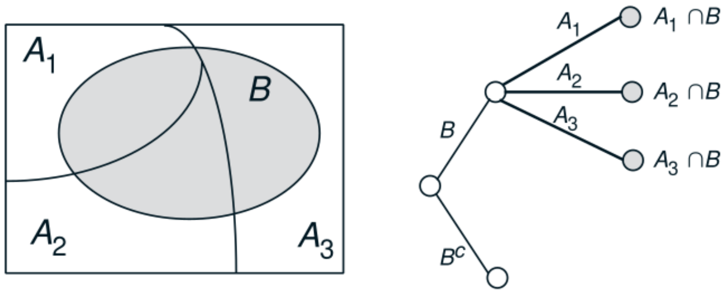
This proves (c). Now we will define some set operations and how they affect probability laws.

First, if two sets A and B are subsets of a sample space Ω, if A and B are **disjoint**, then

$$P(A \cup B) = P(A) + P(B)$$

Total Probability

Let $A_1, A_2, ..., A_n$ be n disjoint sets contained in a sample space Ω which has a probability law P, then: $P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$



The image shows a visualization of three disjoint sets and B, an arbitrary event. See the tree, which gives a sequential description of the problem.

Now, we want to find a way to get the probability of a set B, given that for n disjoint sets $A_1, A_2, ..., A_n$, we have that $B \subset A_i$ for $i = 1, 2, ..., n$ and

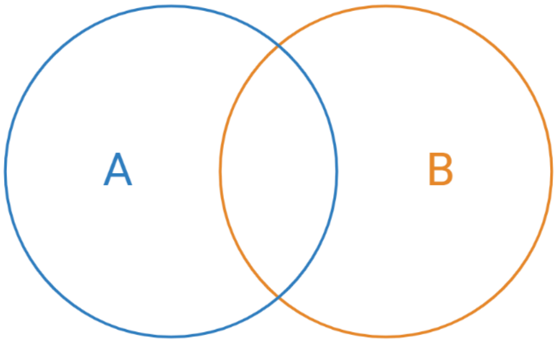
$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup ... \cup (A_n \cap B)$$

Then, taking the probability of B gives

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + ... + P(A_n \cap B)$$

Conditional Probability

To measure how likely is an event A to happen given an event B happened, we can use the notion of sets.



In the diagram we have two joint sets. If we want to know how likely is A to happen given B happened, the solution is to take the ratio between the intersection of A and B and the total probability of B, the equation can be written as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Then $P(A \mid B)$ can be read as "the probability of A given B".

The moral truth

Both $P(A)$ and $P(B)$ represent an "area" in a sample space Ω, where $A, B \subset \Omega$. This means that the equation above is a ratio of two areas.

The notion behind this formula is simple, suppose the intersection $A \cap B$ has "little area", this will cause $P(A \mid B)$ to be smaller, because A and B do not share many points in the sample space. On the other hand, if the area of $A \cap B$ is "bigger", then the probability of A given B should have a greater value, because the fact that B is true makes A have a greater chance of occurring.

Bayes' Rule

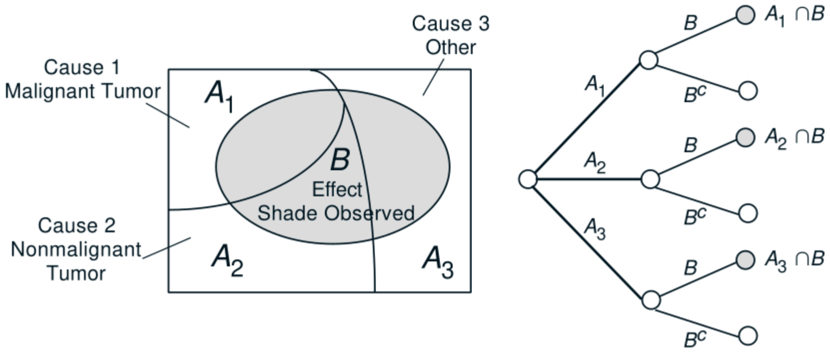
This formula comes from the total probability formula and also from the conditional probability. We will derive the formula directly, suppose B is a combination of intersections between itself and other disjoint sets $A_1, \ldots, A_n$, such that $B = (A_1 \cap B) \cup (A_2 \cap B) \cup \ldots \cup (A_n \cap B)$, then

$$
\begin{aligned}
P(A_i \mid B) &= \frac{P(A_i \cap B)}{P(B)} \\
&= \frac{P(A_i) \cdot P(B \mid A_i)}{P(A_1 \cap B) + P(A_2 \cap B) + ... + P(A_n \cap B)} \\
&= \frac{P(A_i) \cdot P(B \mid A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + ... + P(A_n)P(B|A_n)} \\
&= \frac{P(A_i) \cdot P(B \mid A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}
\end{aligned}
$$

So the final formula is

$$P(A_i \mid B) = \frac{P(A_i) \cdot P(B \mid A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

Here lay some examples of the concepts explained before.



The diagram of a problem which consists of an X-Ray in which a shade is observed. We want to infer what the shade could mean, we have probabilities $P(A_1) = P(\text{"malignant tumor"})$, $P(A_2) = P(\text{"nonmalignant tumor"})$ and $P(A_3) = P(\text{"other cause"})$. Then, for example, finding $P(A_1 \mid B)$, which is the probability of the object to be a malignant tumor given we observed a shade, consists of solving

$$P(A_1 \mid B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(A_1) \cdot P(B|A_1)}{P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + P(A_3) \cdot P(B|A_3)}$$

(The sequential model in the right shows it in other way, the second level of branches in the tree have repeated $B's$, but this does not mean that they have the same probability, instead, in the second branch after A_i, the probability of B is $P(B \mid A_i)$.)

Example 1.14. Alice is taking a probability class and at the end of each week she can be either up-to-date or she may have fallen behind. If she is up-to-date in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.6 (or 0.4, respectively). Alice is (by default) up-to-date when she starts the class. What is the probability that she is up-to-date after three weeks?

Let U_i and B_i be the events that Alice is up-to-date or behind, respectively, after i weeks. According to the total probability theorem, the desired probability $\mathbf{P}(U_3)$ is given by

$$\mathbf{P}(U_3) = \mathbf{P}(U_2)\mathbf{P}(U_3 \mid U_2) + \mathbf{P}(B_2)\mathbf{P}(U_3 \mid B_2) = \mathbf{P}(U_2) \cdot 0.8 + \mathbf{P}(B_2) \cdot 0.4.$$

The probabilities $\mathbf{P}(U_2)$ and $\mathbf{P}(B_2)$ can also be calculated using the total probability theorem:

$$\mathbf{P}(U_2) = \mathbf{P}(U_1)\mathbf{P}(U_2 \mid U_1) + \mathbf{P}(B_1)\mathbf{P}(U_2 \mid B_1) = \mathbf{P}(U_1) \cdot 0.8 + \mathbf{P}(B_1) \cdot 0.4,$$

$$\mathbf{P}(B_2) = \mathbf{P}(U_1)\mathbf{P}(B_2 \mid U_1) + \mathbf{P}(B_1)\mathbf{P}(B_2 \mid B_1) = \mathbf{P}(U_1) \cdot 0.2 + \mathbf{P}(B_1) \cdot 0.6.$$

This problem can be represented by a tree, this tree branches in two for each week that happens. The branch of U events (up-to-date) have probability 0.8, and the branch of B events (left behind) have 0.2 probability. You can find the probability of being up-to-date after the k^{th} week by "backtracking" the tree.

$$\begin{cases} P(U_{k+1}) = P(U_k) \cdot P(U_{k+1} \mid U_k) + P(B_k) \cdot P(U_{k+1} \mid B_k) \\ P(B_{k+1}) = P(U_k) \cdot P(B_{k+1} \mid U_k) + P(B_k) \cdot P(B_{k+1} \mid B_k) \end{cases}$$

Example 1.9. Radar detection. If an aircraft is present in a certain area, a radar correctly registers its presence with probability 0.99. If it is not present, the radar falsely registers an aircraft presence with probability 0.10. We assume that an aircraft is present with probability 0.05. What is the probability of false alarm (a false indication of aircraft presence), and the probability of missed detection (nothing registers, even though an aircraft is present)?

A sequential representation of the sample space is appropriate here, as shown in Fig. 1.8. Let A and B be the events

$$A = \{\text{an aircraft is present}\},$$

$$B = \{\text{the radar registers an aircraft presence}\},$$

and consider also their complements

$$A^c = \{\text{an aircraft is not present}\},$$

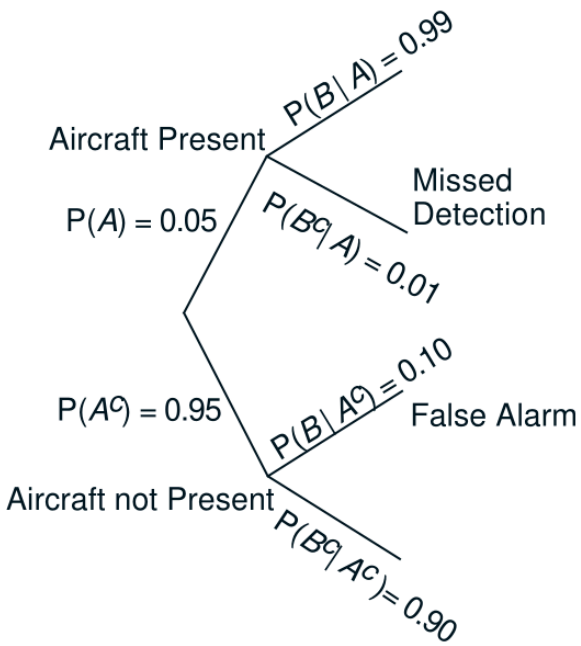
$$B^c = \{\text{the radar does not register an aircraft presence}\}.$$

The given probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.8. Each event of interest corresponds to a leaf of the tree and its probability is equal to the product of the probabilities associated with the branches in a path from the root to the corresponding leaf. The desired probabilities of false alarm and missed detection are

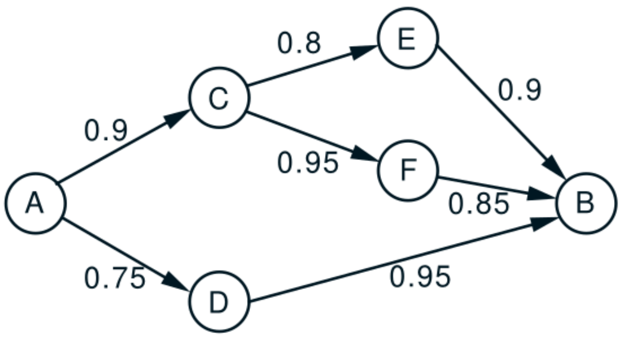
$$\mathbf{P}(\text{false alarm}) = \mathbf{P}(A^c \cap B) = \mathbf{P}(A^c)\mathbf{P}(B \mid A^c) = 0.95 \cdot 0.10 = 0.095,$$

$$\mathbf{P}(\text{missed detection}) = \mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B^c \mid A) = 0.05 \cdot 0.01 = 0.0005.$$

This example of conditional probability can be described by a tree



Now suppose you have some routers A, B, C, D, E, F, G and they are linked as the tree shows, each link or path from a router to the other has a "weight" describing the probability of connection p_i

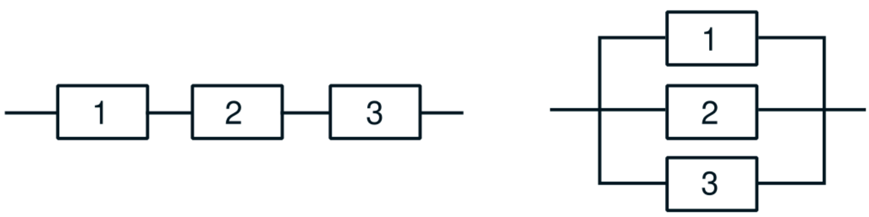


This means that, for example,

$$P(F) = P(A \longrightarrow C) \cdot P(C \longrightarrow F)$$

which becomes $P(F) = (0.9)(0.95) = 0.855$.

Now suppose we can connect nodes in series or parallel as shown below



We represent a **series network connection** between a router A which has to find a route through other routers $C, D, E, F, G, ...$ to find a router B, by the product of probabilities. Because it is a sequential model, if one router fails, then everything should be affected

$$P(A \longrightarrow B)_{series} = p_1 p_2 ... p_n$$

here $p_1, p_2, ..., p_n$ are the probabilities of each link in the path from A to B.

To represent a **parallel network connection**, each link has a $1 - p_i$ probability to fail, we take the probability of failure because for the whole network to fail, all routers must fail

$$P[\neg(A \longrightarrow B)] = (1 - p_1)(1 - p_2)...(1 - p_n)$$

Then, we know the probability of a connection from A to B for a parallel connection to be

$$P(A \longrightarrow B)_{parallel} = 1 - (1 - p_1)(1 - p_2)...(1 - p_n)$$
