[STAT-03] Introduction to probability

Miguel-Angel Canela

Associate Professor, IESE Business School

Why probability?

When we perform a statistical analysis, we wish to draw conclusions that apply, not only to the **sample** to which the data refer, but to a **population** from which the sample has been extracted. This is called **statistical inference**. For instance, market research aims at the population of all potential customers, though the data only cover a sample of actual customers.

Inference has to be performed with care, since statistical results are always based on partial information. So, these results are usually reported with an indication of the extent to which we can trust them. The probability language allows us to manage this.

There are certain standards that our statistical analysis should satisfy, if we want our paper to be published. One of them is that the results reported must be **significant**. Roughly speaking, this means that we should be, at least, 95% sure of our conclusions. In technical terms, that the probability of obtaining results such as those reported would be less than 0.05 if our hypothesis were not valid. But, it is enough for us to have an idea of the probability of an event as a numerical measure of how likely it is to happen, and to know the rules for operating with probability values.

Different approaches to probability

There have been various attempts to define the probability in a consistent way. Some of them are based on philosophical or psychological ideas about how humans deal with uncertainty, like **subjective probability**. Leaving these apart (not meaning that they are not interesting), there are two main perspectives:

- In the **frequentist** approach, the probability of an event is the limit of the proportion with which it occurs, as the number of observations increases. For instance, when saying that the probability of a newborn being male is 52% (a realistic figure), we understand that, in a large number of births, the outcome of (approximately) 52% of them will be a male child. So, in the frequentist approach, probabilities are understood as related to **large numbers**.
- In the **Bayesian** approach, the probability is a numerical measure of our expectations about the occurrence of an event. So, it is related to the information available. With new information, the probability changes. The Bayesian scientist starts with a **prior** probability model. After collecting the data, the model is updated to a **posterior** model.

You should not care about this in this course. We just take the probability of an event as an assessment of how likely that event is to occur. What matters is the set of rules under which we assign probabilities to events.

Formal definition

The formal definition of the probability was not easily achieved. Indeed, it took about two hundred years, from the first calculations related to gambling strategies, to the **probability axioms**, formulated by the Russian mathematician AN Kolmogorov, used today as a formal definition.

To account for the mathematical standards, I use the language of set theory, taking the **events** whose probabilities are discussed as if they were sets. More specifically, the events are subsets of a set S called the **sample space** (to get intuition, think on S as the set of all possible outcomes of an observation or experiment). This allows us to consider as events both the **empty set** $\emptyset$, which stands for the impossible, and S, which stands for all possible results. We can also combine events A and B, getting the **union** $A \cup B$ (A or B), the **intersection** AB (A and B) and the **complementary event** A^c (not A).

The set of events satisfies the following axioms:

[E1] The empty set $\emptyset$ is an event.

[E2] If A is an event, then the complement A^c is also an event.

[E3] Given a sequence of events A_1, A_2, ..., A_n, ..., the union $\bigcup_{n=1}^{\infty} A_n$ is also an event.

A **probability** is a function defined on the events, satisfying the axioms:

[P1] For every event A, $\mathrm{p}[A] \geq 0$.

[P2] $\mathrm{p}[S] = 1$.

[P3] Given a sequence of pairwise disjoint events A_1, A_2, ..., A_n, ...,

$$\mathrm{p}\left[\bigcup_{n=1}^{\infty} A_n\right] = \sum_{n=1}^{\infty} \mathrm{p}[A_n].$$

The sample space must be understood as a mathematical entity that allows the use of the language of set theory. But, as far as we can manage the events and their probabilities through sound mathematics, the actual sample space is irrelevant. I illustrate this with the following example.

Example 1. We want a model for the probabilities associated to two basic events, even (A) and odd (B), in the outcome of a regular die. Of course, we can take

$$S = \{1, 2, 3, 4, 5, 6\}, \qquad A = \{2, 4, 6\}, \qquad B = \{1, 3, 5\}.$$

We have here four events $\emptyset$, S, A and B, with $\mathrm{p}[\emptyset] = 0$, $\mathrm{p}[S] = 1$, $\mathrm{p}[A] = \mathrm{p}[B] = 0.5$. But we can also consider

$$S = \{e, o\}, \qquad A = \{e\}, \qquad B = \{o\}$$

(e stands for even and o for odd), with the same probabilities. There is no difference between the two models, as far as we do not want to consider other events.

Other formulas

Some consequences of these axioms, which can be proved without much pain, are:

- $\mathrm{p}[\emptyset] = 0$.
- $\mathrm{p}[A^c] = 1 - \mathrm{p}[A]$.
- If $A \subseteq B$, then $\mathrm{p}[A] \leq \mathrm{p}[B]$.
- $0 \leq \mathrm{p}[A] \leq 1$.
- $\mathrm{p}[A \cup B] = \mathrm{p}[A] + \mathrm{p}[B] - \mathrm{p}[AB]$.

Example 2. A patient arrives to the doctor's office with a sore throat and low-grade fever. The doctor is sure that the patient has either a bacterial, or a viral infection, or both. He attributes probabilities 0.7 to the bacterial and 0.4 to the viral infection. What is the probability that a patient has both?

Calling B and V the events associated to the two infection types, we have $B \cup V = S$, since the doctor is sure that these are the only possible causes of the symptoms. Then

$$\mathrm{p}[BV] = \mathrm{p}[B] + \mathrm{p}[V] - \mathrm{p}[B \cup V] = 0.7 + 0.4 - 1 = 0.1. \; \square$$

The formula of the probability of the union of three events,

$$\mathrm{p}\big[A \cup B \cup C\big] = \mathrm{p}[A] + \mathrm{p}[B] + \mathrm{p}[C] - \mathrm{p}[AB] - \mathrm{p}[AC] - \mathrm{p}[BC] + \mathrm{p}[ABC],$$

is proved by writing the union of three events as $A \cup B \cup C = \big(A \cup B\big) \cup C$ and applying the formula for the union of two events. The extension to n events,

$$\mathrm{p}\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} \mathrm{p}[A_i] - \sum_{i<j} \mathrm{p}\big[A_i A_j\big] + \sum_{i<j<k} \mathrm{p}\big[A_i A_j A_k\big] + \cdots + (-1)^{(n+1)} \mathrm{p}\big[A_1 A_2 \cdots A_n\big],$$

is not hard to get, although the use of subscripts may scare you.

A comment on zero probability. We always have $\mathrm{p}[S] = 1$ and $\mathrm{p}[\emptyset] = 0$, but there may be other pairs of complementary events with this property. This is not easily understood, because we intuitively look at the sample space as a finite set of pieces. But, in practical statistical analysis, we deal with situations more complex than that. For instance, an event that contains a single value of a continuous variable has probability zero. This does not contradict axiom [P3], because an interval of the real line cannot be expressed as a sequence of points (it contains too many points). Mathematical consistency forces us to accept a paradoxical fact, that every time that the variable takes a particular value, something with probability zero occurs.

Statistical independence

Statistical independence is one of the central notions of probability. In academic research, most of the statistical analyses are performed to provide evidence for rejecting a statistical independence statement. Although this statement is usually formulated in terms of variables, this lecture only considers events, leaving variables for later.

We say that two events A and B are statistically independent when

$$\mathrm{p}\big[AB\big] = \mathrm{p}[A]\,\mathrm{p}[B].$$

Note that an event of probability zero is statistically independent of any other event. Why am I so fastidious putting the adjective "statistical" before the noun "independence", instead of talking plainly about independence? To emphasize that independence is not an intrinsic property of the two events, since it depends on the probability. In many problems, we can consider two different probabilities, and two events may be independent with respect to one probability but not with respect to the other one. In particular, they may be conditionally independent, but not independent, as we will discuss in the next lecture.

The independence of a collection of events is defined as follows. The events $A_1, \ldots, A_n$ are said to be statistically independent when, for every subcollection $A_{i_1}, \ldots, A_{i_k}$,

$$\mathrm{p}\big[A_{i_1} \cdots A_{i_k}\big] = \mathrm{p}[A_{i_1}] \cdots \mathrm{p}[A_{i_k}].$$

It must be remarked that pairwise independence does not imply independence. This matters, because, as you will realize by yourself, we intuitively look at dependence in terms of pairwise relationships. The fact that pairwise dependence is usually assessed through a correlation coefficient facilitates this. Here follows a simple counterexample.

Example 3. Consider an experiment with four outcomes s_1, s_2, s_3 and s_4, with probability $1/4$, and the events

$$A = \{s_1, s_2\}, \qquad B = \{s_1, s_3\}, \qquad C = \{s_1, s_4\}.$$

Then A, B and C are pairwise statistically independent, but not statistically independent. First, note that $\mathrm{p}[A] = \mathrm{p}[B] = \mathrm{p}[C] = 1/2$. Now, $AB = AC = BC = ABC = \{s_1\}$, hence

$$\mathrm{p}\big[AB\big] = \mathrm{p}\big[AC\big] = \mathrm{p}\big[BC\big] = \mathrm{p}\big[ABC\big] = 1/4,$$

justifying my assertion.