[STAT-02] Probability calculus

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, those 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 **statistically 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 the hypothesis that the paper intends to prove 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 do not need worry about that in this course, since we just take the probability of an event as an assessment of how likely that event is to occur. What matters here 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 discussions about gambling strategies, to the **probability axioms**, formulated

by the Russian mathematician AN Kolmogorov, which are used today as a formal definition.

To account for the mathematical standards, I use here 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 infection, 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 emphasise 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 later in this 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.

Conditional probability

Let A and B be events, with $\mathrm{p}[B] \neq 0$. The probability of A **conditional** to B is defined as

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

This definition only makes sense when the conditioning event has non-zero probability. For two nonzero probability events A and B, statistical independence is equivalent to $\mathrm{p}[A] = \mathrm{p}\big[A|B\big]$. This provides an intuitive definition of statistical independence: A and B are independent when knowing that B occurred does not change the probability of A.

Example 4. What is the probability that the sum of the outcomes of two dice is less than 8, if we know that it is odd? Let A the event that the sum is less than 8 and B the event that it is odd. The numerator and the denominator in the definition above can be easily calculated by counting cases and dividing by 36. Cancelling out 36, we get

$$\mathrm{p}\big[A|B\big] = \frac{2+4+6}{2+4+6+4+2} = \frac{2}{3}\,. \;\square$$

Conditional probability comes easily when:

- Probabilities are regarded as based on the information available. A frequent question is: what is the probability of A, *knowing that* B has occurred? The answer would be the probability of A conditional to B, which I denote here by $\mathrm{p}[A|B]$.
- We focus on a subpopulation. Take, for instance, the event of a newborn being male. If we consider the probability of this event, but conditional to the mother being older than 30, we are restricting the analysis to a subpopulation of the population of births.

Note that, since observational studies are usually restricted to a certain subpopulation (companies big enough, male executives, working wives, etc), considering a probability as conditional is just a matter of convenience. We do it when it is practical to do so.

A conditional probability is itself a probability. That is, if we define $\mathrm{p}_B[A] = \mathrm{p}\big[A|B\big]$, we get a probability p_B on the same set of events. It is easy to understand this probability: (a) any event that does not meet B has probability zero, and (b) for the rest, the probability of A is the probability of the intersection AB, rescaled by dividing by $\mathrm{p}[B]$.

The multiplication rule

From the definition of the conditional probability, we get directly the **multiplication rule** (note that the roles of A and B are interchangeable),

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

The multiplication rule is extended to n events,

$$\mathrm{p}\big[A_1 \cdots A_n\big] = \mathrm{p}[A_1]\,\mathrm{p}\big[A_2|A_1\big]\,\mathrm{p}\big[A_3|A_1A_2\big] \cdots \mathrm{p}\big[A_n|A_1A_2 \cdots A_{n-1}\big],$$

by applying the rule successively to p, p_{A_1}, $\mathrm{p}_{A_1A_2}$, etc.

Example 5. Imagine a box containing 5 red balls and 5 blue balls. If we draw three balls at random, what is the probability that the first two are red and the last one is blue?

For the first ball, the probability of being red is 5/10, for the second ball, given that the first ball is red, the probability of being red is 4/9. Finally, for the third ball, given that the first two have been red, the probability of blue is 5/8. Therefore, the probability sought is the product, 0.139.

Bayes formula

I present in this paragraph two important formulas. First, for a partition $B_1, \ldots, B_k$ of the sample space, we can write $A = (AB_1) \cup \cdots \cup (AB_k)$ and apply axiom [P3], getting

$$\mathrm{p}[A] = \mathrm{p}[B_1]\,\mathrm{p}[A|B_1] + \cdots + \mathrm{p}[B_k]\,\mathrm{p}[A|B_k].$$

This is the **formula of total probability**. Note that $\mathrm{p}[A]$ appears as a weighted average of the probabilities of A in the k cases defined by the partition. The weights are the probabilities of these cases.

The **Bayes formula**,

$$\mathrm{p}[B|A] = \frac{\mathrm{p}[A|B]\,\mathrm{p}[B]}{\mathrm{p}[A]}\,,$$

is the cornerstone of Bayesian statistics. It follows directly from the definition of the conditional probability. Combining Bayes formula with the formula of total probability, we get

$$\mathrm{p}[B_i|A] = \frac{\mathrm{p}[A|B_i]\,\mathrm{p}[B_i]}{\mathrm{p}[B_1]\,\mathrm{p}[A|B_1] + \cdots + \mathrm{p}[B_k]\,\mathrm{p}[A|B_k]}\,.$$

Example 6. An insurance company has three types of costumers: high, medium and low risk. 20% of the customers have high risk, 30% have medium risk, and 50% low risk. The probability that a customer has at least one accident in the current year is 0.25 for high, 0.16 for medium and 0.10 for low risk. What is the probability that a customer has high risk, given that he/she has had at least one accident during the current year?

I denote by R_1, R_2 and R_3 the three risk groups, and by A having at least one accident. Then,

$$\mathrm{p}[R_1] = 0.2, \quad \mathrm{p}[R_2] = 0.3, \quad \mathrm{p}[R_3] = 0.5, \quad \mathrm{p}\big[A|R_1\big] = 0.25, \quad \mathrm{p}\big[A|R_2\big] = 0.16, \quad \mathrm{p}\big[A|R_3\big] = 0.10.$$

Now, Bayes formula gives

$$\mathrm{p}\big[R_1|A\big] = \frac{0.2 \times 0.25}{0.2 \times 0.25 + 0.3 \times 0.16 + 0.5 \times 0.10} = 0.338.$$

Conditional independence

As mentioned above, a conditional probability is a probability. When we consider independence with respect to this particular probability, we call it **conditional independence**. So, conditional to C, A and B are (statistically) independent when

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

The extension to more than two events is done in the obvious way. The following equivalent statements provide different views of conditional independence:

- A and B are independent conditional to C.
- $\mathrm{p}\big[A|BC\big] = \mathrm{p}\big[A|C\big]$.
- AC and BC are independent.

Two events can be conditionally independent without being independent. This may seem, at first sight, an excuse for the professor to introduce an exotic counterexample, but it is not. Many theoretical issues discussed in managerial science papers are concerned with this distinction, which can be seen as a particular case of a more general topic, the **Simpson paradox**. There are many versions of this paradox, always relying on the same fact: conditioning changes, sometimes dramatically, the probability distribution.

I restrict the actual discussion to the simplest issue, the distinction between conditional and unconditional independence, illustrating it with a challenging example, which shows that we can take two tosses of one coin as either dependent or independent observations.

Example 7. Being used to take successive coin tosses as statistically independent in elementary probability exercises, we don't bother mentioning it. But it is not so simple, since what we really have is independence conditional to the coin being specified. Let me consider an experiment involving two coins, one with one head and one tail and the other with two heads. I first choose at random the coin, and then I toss twice the selected coin.

I denote by F the choice of the fair coin, and by H_1 and H_2 getting head in the first and second toss, respectively. Applying the formula of total probability to the partition formed by F and F^c, we get

$$\mathrm{p}[H_1] = \mathrm{p}[H_2] = \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot 1 = \frac{3}{4}\,, \qquad \mathrm{p}\big[H_1H_2\big] = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot 1\cdot 1 = \frac{5}{8}\,,$$

showing that H_1 and H_2 are not independent. Note that

$$\mathrm{p}\big[H_2|H_1\big] = \frac{\mathrm{p}\big[H_1H_2\big]}{\mathrm{p}[H_1]} = \frac{5}{6}\,.$$

A Bayesian statistician will tell you the story in the following terms. A priori, the probability of getting head in the second row is 3/4. After observing the first toss (a posteriori), the probability is 5/6, higher because I got head in the first row (if I had gotten tail, it would be lower, 1/2).

Homework

A. In a town of $n+1$ inhabitants, a person tells a rumor to a second person, who in turn repeats it to a third person, etc. At each step, the recipient of the rumor is chosen at random among the other n persons. Find the probability of the rumor being told exactly r times (including the first person telling it to the second person):

(a) Before returning to the originator.

(b) Without being repeated to any person.

B. There are many versions and extensions of the *Monty Hall problem*, all of them descendants of a probability classic, the *Bertrand's box paradox*. Monty Hall was the host on a TV classic called *Let's Make a Deal*. Monty gave the player the choice of three doors: behind one door was a car and, behind the other two doors, goats. The player picked a door, say No. 1, and Monty, who knew what was behind the doors, opened another door, say No. 3, which had a goat. He then said to the player,"Do you want to pick door No. 2?"

Would you switch the initial choice?