Probability and statistics

Statistics and probability calculations are two mathematical disciplines of crucial importance not only for scientific thought but also because they can help us, in everyday life, to form an opinion or to make an "informed" decision.

To begin to understand what probability and statistics consist of, we give an overview of the relationships that exist between them, highlighting the differences between the two.

	Probability	**Statistics**
Scientific area	Maths	Maths
How can this help us?	to judge / choose rationally on a quantitative basis in an uncertain context	to judge / choose rationally on a quantitative basis in an uncertain context
When and how this emerged as a modern discipline	in the XVII century, from the study of problems related to gambling	in the XVII-XVIII centuries, for the information needs of the national states
Mainly used for	predicting the future	interpreting the past, learning from it, interpreting the present
Types of reasoning this supports?	deductive	inductive
Type of discipline	theoretical; it can be developed starting from a few axioms	practical; helps us to hypothesize laws by observing regularity in the data
Allows us to	evaluate the consequences of the laws of a certain ideal world	measure how our world deviates from an ideal world

One way of understanding the differences between these two disciplines is to consider the possible thought processes of a mathematician who first meets the game of dice [2].

- If the mathematician was a probability expert, in jargon a “probabilist", observing a dice he would think: A dice with 6 faces? Each face has the same odds of landing. Now, assuming that every face has a probability of 1/6, I can understand what my chances are of winning by betting on a combination in a sequence of throws.

- A statistician, on the other hand, would see the dice and think: These dice may seem Ok, but how do I know they are not rigged? I will look them for a while and keep track of how often each number comes out. Then I can decide whether my observations confirm the assumption that the sides have equal probability; when I am sure, I will ask a probabilist to advise me how to play .

Probability

The probability is presented in many variations and associated with many different terms, such as "expectation" , "case", "possibility", “chance”, "odds" (of a competitor in a race), etc. Probability is a useful notion when there is no *certainty*, that is, in the processes that are intrinsically *random* or that we consider as such because we do not know well the laws to which they obey.

The world is full of uncertainty

Who will win the championship?

Will the stock index rise or fall today?

What will the weather be today?

Will I like this movie?

We - we personally, or an 'intelligent' system that we use to obtain a forecast or advice - are constantly interacting with the world, and we need to get a feel for it: what is it that we see or hear and how it might behave or to evolve?

It is rare that we can prove that something is true, but we still wonder how likely a certain event is or what is the most reasonable explanation of a certain fact.

Although we commonly speak of scientific "certainties", experts know well that there are very few certainties demonstrated with the same rigor of a theorem of mathematics.

For example, for experimental physicists two sources of uncertainty are: the so-called "measurement errors", which leave us uncertain about the value of the observed data, and the difficulty of identifying the theory (most) compatible with the observed phenomena. [9]

The scientist's difficulties also depend on the fact that, as a first approximation, there are two types of laws:

- probabilistic laws; a simple example is the coin toss; a more complex example is that of genetics; in these fields, observable events are not a pure logical consequence of the theory
- deterministic laws; the "classic" example is provided by Newton's laws (*classical mechanics*), which are valid only at a certain scale of observation and with some limitations.

A definition and two axioms

Among the probability definitions that we encountered, this seemed to us quite simple:

> The probability of an event is how many times we expect, *in the long term*, that this *event* occurs as a result of a *random process*, compared to all the other possible results.

Even if, as we will see later, this definition is a bit ambiguous, we already got some fixed points: a) a probability is a *ratio* ; b) its value is a *number* between 0 and 1; sometimes this value is expressed as a percentage or as a fraction in which the numerator represents the number of "successes" expected with respect to the total number of tests. The two extremes, zero and one, represent borderline cases, because in reality they correspond to a certainty.

The process discussed in the definition could consist in the repeated throwing of one or more dice, or in playing heads (H) or tails (T); in the latter case, if you were to play 3 times, an event could consist of the sequence [HHT]; the remaining events correspond to the other 8 possible combinations: [HHH], [HTH], [THH], etc.

And with this we are already able to introduce two of the three *basic axioms* of probability: [6]

> The *first axiom* , also called the axiom *of positivity*, tells us that the probability of a generic event *E* is a real number greater than or equal to 0; in symbols:
>
> P(E) >= 0
>
> The *second axiom* , also called the *axiom of certainty*, tells us that the probability of the *certain event I* is 1; in symbols:

$$P(I) = 1$$

an alternative formulation is this: the probability of the entire *sample space*, symbolized by Ω, i.e. the sum of the probabilities of all possible alternative events, is 1; in symbols:

$$P(\Omega) = 1$$

Coming back to the definition of probability, what can leave you perplexed is the expression "in the long term"; what does it mean? Someone replaces it with the expression "in an infinite number or experiments"; but this makes the definition non-operational, i.e. impossible to use to determine the probability of an event in an experimental way.

An alternative definition of probability, which a theoretical probabilist would perhaps like more, is the following:

> The probability of an event is the ratio between the number of favorable cases and the number of possible cases, all assumed *equally possible*.

This definition also leaves us a bit unsatisfied, because it is not easy to identify application domains in which all cases are equally possible. Usually these are rather artificial domains, like those of gambling games.

On the other hand, a direct definition of probability is not necessary; we have already seen that the calculation of probabilities is a theoretical discipline; that it can be developed starting from a few axioms: are these axioms and all the theorems that we can derive from them that constitute the true (indirect) definition of probability.

The third axiom

At this point we must introduce the third axiom, which is slightly more complicated than the others. And to do that, let's go over the terminology we already met and introduce some more: [7]

- *random process*: it is a natural phenomenon or an experiment of which we can observe a series of results (or outputs) that we are not able to predict
- *event*: it is a specific result or set of results within a random process
- *simple event*: it is an event that contains only one result, for example the throwing of a dice
- *compound event*: it is an event that contains more results, for example a sequence of dice throws
- *space sample* (or *sampling space*): it is the set of all possible events in a random process
- *incompatible events*: events that cannot occur simultaneously; for example two simple dice throwing events, in one of which the 3 comes out and in the other the 4.

And here is the *third axiom*, also called the *axiom of union* :

> If A and B are incompatible events, then the probability of the *union* of the two events (i.e., that either A or B occurs) is equal to the sum of the probabilities of the individual events. In symbols:
>
> $$P(A \cup B) = P(A) + P(B)$$
>
> where we can read "A ∪ B" as "A union B”

Example: if we roll a dice only once, the probability of "exiting 3 or 4" is the sum of the probability of "exiting 3" and the probability of "exiting 4". It seems too obvious that different faces of a dice correspond to incompatible events; other types of experiment, such as the extraction of a card from a deck, provide more significant examples; we leave the reader to identify, among the following pairs of events, the only one that includes incompatible events: [8]

a) extracting a card from a deck: E1 - "a sword card comes out", E2 - "a cup card comes out"

b) playing roulette: E1 - "comes out a number between 10 and 30", E2 - "red comes out"

c) extracting a card from a deck: E1 - "a figure comes out", E2 - "a coin card comes out".

Reasoning with conditional probability

Independent events and compound probability

Consider the sequence of two events of this example:

throwing a coin 2 times: A - "head comes out at the first throw", B - "tail comes out at the second throw"

we probably already know that the two events are *independent*, that is, that the result of the first throw does not influence that of the second one; believing the opposite is the typical *fallacy of the gambler*.

The *compound probability theorem* states that

In the case of independent events, the probability of a compound event is equal to the product of the probabilities of the component events. In symbols:
$P(A \cap B) = P(A) * P(B)$

where we can read "$A \cap B$" as "A intersection B"

Using this theorem we can compute the compound probability for our example:

$P(A \cap B) = P(A) * P(B) = 1/2 * 1/2 = 1/4$

given that, in the case of an ideal, symmetrical coin, the probability that each of the two faces comes out is 1/2.

Dependent events and conditional probability

Sometimes it happens that we have to calculate the probability of the occurrence of an event, knowing that another event, connected to the first one for logical reasons or for temporal succession, has already occurred or cannot not occur.

With the notation $P(A \mid B)$ we indicate the *conditional probability of* A *given* B, defined as the probability that event A has if event B is certain.

From the three basic axioms of probability calculations, the *conditional probability theorem* can be derived , which is expressed by the following formula

$P(A \mid B) = P(A \cap B) / P(B)$

it states that

The conditional probability of an event A given an event B, which is possible, is equal to the ratio of the compound probability of the two events and the absolute probability of B.

The conditional probability theorem is useful in the case of events that depend on each other, while if we tried to calculate the conditional probability in the case of independent events, we would realize that it coincides with the absolute probability. Conversely, the two equalities

$P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$

jointly constitute a possible definition of the reciprocal *independence* of events A and B.

Now consider these two pairs of events :

throwing a dice: A - "the 5 comes out" , B - "an odd number comes out "

throwing a dice: A - "the 5 comes out" , B - "an even number comes out"

since 5 is one of the three odd faces, in the first case

$P(A \mid B) = 1/3$, or the conditional probability of A given B is a third;

given that among the even numbers the 5 does not appear, in the second case

$P(A \mid B) = 0$, or the conditional probability of A given B is zero

while in both cases the absolute probability of A is 1/6; therefore, in both cases the two events of the couple are *dependent*.

Statistics

As we have seen in the previous unit, the formalization of the social sciences, such as psychology and economics, is more problematic than that of the so-called "hard" sciences, such as chemistry and physics; however, even in the "soft" disciplines, the use of mathematics is not given up.

To deal with the aspects and models of reality that are faced in the social sciences, the branch of mathematics that is most developed is statistics; it is difficult to clearly distinguish it from the theory of probability, given that also in it the reduction of the uncertainty is one of the central objectives.

Statistics is a fairly broad and varied branch of mathematics, which was developed primarily to support data interpretation, search for regularity and trends in the data, and formulation of forecasts in disparate fields of application, such as, for example

- genetics, demography
- environmental sciences, resource analysis
- econometrics, market research
- epidemiology, experimentation with new drugs
- sociology, experimental psychology
- computational linguistics.

A tale of widows and orphans

Statistics, as we know it today, was born 2-3 centuries ago, driven also by the needs of the offices that manage life insurance. An interesting story in this regard was told in a conference of academic actuaries in 1972 [11] and is also taken up and summarized in [3] and [4]. Here we make a further summary from [3].

It seems that the Scots are the creators of the modern business of life insurance, a business that was born from the friendship of two ministers of the church, Robert Wallace and Alexander Webster, lovers of mathematics and concerned about the fate of widows and of orphans of their colleagues; in fact, from the moment when the reformed church of Scotland allowed its ministers to marry, at the death of these widows and orphans often fell into severe misery.

Between one drink and another, because they were known drunkards, Wallace and Webster in 1744 developed, and had approved by law, a plan to create a perpetual pension fund; the most difficult problem was to understand how much every minister should periodically pay to this fund. To understand this, they gathered from various sources, such as the parish registers, data useful for estimating the life expectancy of the Scottish ministers, as well as data on average rates of investment income.

The work of Wallace and Webster led to the elaboration of the first *mortality tables*; in this they used actuarial tables published half a century before by the English mathematician Edmond Halley, the advice of a professor from the University of Edinburgh and the recent advances in the calculation of probabilities, among which the most notable was the formulation of the *law of large numbers* by the Swiss mathematician Jakob Bernouilli.

All this allowed the two to calculate how many ministers would be alive and how many would die on *average* each year, how many years *on average* widows would survive and with how many children, etc. and so on; in the end they decided that the pension fund would be supported if every minister had paid at least 10 pounds a year. They also estimated that from then to 1765 the fund would have accumulated a capital of 58.438 pounds; their prediction turned out to be surprisingly precise: that year the fund's capital reached just that figure, less than a pound!

The law of large numbers

It is a widespread observation among the experimenters that, when dealing with random phenomena, the relative frequency of an event, i.e. the number of times that event occurs with respect to the total number of tests, tends to *stabilize* as the number of tests increases; this is called *the empirical law of the chance*.

In the early 1700s the empirical law of the chance was reformulated by Jakob Bernouilli, who renamed it as the *law of large numbers*, in the following terms:

> If *P* is the *probability* of success of an event *E*, that is the probability of the occurrence of *E* in a test, then the *relative frequency* of successes in *N* independent trials tends to *P* when *N* tends to infinity .

The law of large numbers, also called *Bernouilli's Theorem*, is important because it clearly relates two concepts to each other

- the theoretical *probability* of an event, which is the central concept of probability theory
- the *relative frequency* of an event, which is the central concept of statistics.

This law is of no use to predict the success or otherwise of an event from time to time (even if some try to do it, at their own risk!); it says instead that, even if it is difficult to predict the outcome of a single event, for example, the *life time of* a radioactive nucleus, it can be easy to estimate the average life span of *N* radioactive nuclei: the greater the number *N*, the better the approximation.

A simple quiz

We have just mentioned one of the fundamental notions of statistics, the "mean"; let's try to do a little test on it. To get the average number of children per family in a village, a teacher counted the total number of children. He then divided by 50 because there were 50 families. The average number of children per family was 2.2. Which of the following statements is certainly true? [1]

(a) Half of the families in the village have more than two children

(b) More families in the village have 3 children than 2 children

(c) There are a total of 110 children in the village

(d) There are 2.2 children in the village for each adult

(e) The most common number of children in a family is 2

(f) None of the above statements is true.

Certainly you will be able to find the right answer on your own; but how long?

Basic terms and concepts of statistics

In statistics we use to distinguish two main strands:

- *descriptive statistics*; collects the data, summarizes them and comments on them to describe the characteristics of a *statistical population*; a population can be a group of people, animals, plants, etc.; the elements of the population are called *statistical units*
- the *inferential statistics*; uses data to induce rules and make predictions.

With reference to the set of numerical values of a certain property of a *population*, statistics considers certain derived quantities or *parameters*; the two most important and known parameters are:

- the *average value*, also called the *arithmetic mean*; is the ratio between the sum of the values of the considered property and the number of such values, that is of the statistical units to which the values refer:

  $(v_1 + v_2 + \ldots + v_n)$

  in our example, the population is made up of the 50 families of the village, each family is a unit, the property that we consider is the number of children, the sum of the values of this property is 110 and the average value is 110/50 = 2.2

- the *median value* , also called the *median*; the median is the central value, such that half of the population has a value lower than it for the property considered, and the other half of the population a higher value; in the example on the families of the village, we do not have enough information to identify the median.

Sampling and statistical inference

Censuses and sampling

In principle it is possible to derive all the parameters of interest of a statistical population by carrying out a *census survey*, which involves the entire population. The most typical census survey is the *census of the population* of a region: news of censuses, often limited to some categories of people, are found in almost all the ancient peoples.

In consideration of the complexity, duration and cost of census surveys, the practice of *sample surveys*, i.e. carried out on a limited sample of the population, has become established in modern times; while the first ones directly supply the parameters of interest, in the case of sample surveys, such as the electoral polls and the market surveys, those parameters are estimated .

In many cases a well-designed sample survey can provide more accurate results than a census survey carried out with limited means and preparation. In Italy the sampling method was introduced in the practice of ISTAT (the Italian National Institute of Statistics) in 1925.

Sophisticated statistical techniques are used

- to *choose* a representative sample; there is in fact a wide variety of sampling methods and not necessarily a "random" sampling is the best
- to *infer* the parameters of the entire population from the parameters calculated on the sample and to estimate the magnitude of the consequent approximation errors.

Statistical inference

Statistical inference is a set of methods used to try to draw conclusions about a population on the basis of information obtained from a sample. It can be carried out autonomously, to study without "preconceptions" the characteristics of an entire population or be carried out in an auxiliary way, to build *scientific evidence* on a research hypothesis, especially when studying phenomena with a high variability such as biological phenomena .

In the second case, the method, a variant of the more general *inductive cognitive cycle*, involves the following steps

- formulation of the hypothesis.
- evaluation of the probability of obtaining certain results in the population if the hypothesis was true
- extraction of a part of the population (sampling)
- computation of sample statistics (example: average values on the sample)

- estimation of the parameters in the population, based on the results provided by the sample (inference).

The Bayesian approach to learning

Bayes Theorem

The *Bayes theorem*, also known as *the probability of causes theorem*, or the *probabilistic causation theorem*, can be considered the fundamental theorem of the reasoning form called *abduction*. Proposed by Rev. Thomas Bayes in the '700, it is used to calculate the probability of an event A as the triggering cause of a verified event B. It "reverses" the conditional probability passing from the "a posteriori" one to the "a priori" one. [13]

The statement of *Bayes' theorem* is not particularly complex:

> The conditional probability of an event A given an event B is equal to the conditional probability of B compared to A multiplied by the ratio of the absolute probability of A and B.
>
> P(A | B) = P(B | A) * P(A)/P(B)

Even its proof is not difficult, but we omit it as in the case of the other theorems.

In fact, more that the *Bayes' theorem* as explained above, we are interested in its extension to the case where we know that events A_1, A_2, ... A_n, being completely disjoint, fully cover the sample space *A* and we know the absolute probabilities of each of them; in this case, the extended Bayes' theorem allows us to calculate the probability of each event A_i as a cause of the verified event *B*.

Instead of showing the formulation of the extended version of the theorem, we describe with an example the typical problem that it allows to solve. [13]

> A company produces a computer component in three different plants, which we will call P1 , P2 and P3. The components produced may or may not be defective. We know that the company produces 30% of the total components in the P1 plant , 25% in the P2 plant and the remaining 45% in the P3 plant. Statistical surveys conducted by the company confirm that 2% of the components produced in the first plant is defective, in the second plant the defect occurs with an incidence of 1.8%, while the third plant only produces 1.33% of defective components. A customer orders a component and receives a defective component. Calculate the probability that the component received comes from the second plant.

The solution is: the defective component comes with a probability of 27.3% from plant P2. We do not give the demonstration, also because we have not exposed the extended version of the Bayes theorem on which it can be based; on the other hand, we were only interested in giving an idea of its possible use.

Bayesian inference and machine learning

Bayes theorem is considered a cornerstone in the construction of models of how one can learn from experience, making good use of available data; then it is not surprising the fact that the so-called *Bayesian inference*, a methodology inspired by it, is so popular in the field of *machine learning* .

The typical application context is that of *complex systems*, systems in which several *agents* interact in a non-linear fashion, such that the effects do not depend on the causes in a simple, proportional way. When they are not too complex, we try to represent their behavior through *parametric* models; the definition of a model requires the identification of the values of the *parameters* that appear in it; there are often theoretical reasons to assume that these values follow well-studied *statistical distributions*, such as the known *Bernouilli*

distribution for *discrete* value events (dice throwing) or the *Gauss distribution* (the typical Gaussian curve, which resembles a bell) for events with continuous values.

Bayesian inference is a statistical inference method in which the Bayes' theorem is used to update the probability of a hypothesis , i.e. the parameters of the model that represents it, as more evidence or information becomes available. [15]

In Bayesian inference, unlike in more traditional approaches, the model parameters are learned in the form of *probability distributions*, which makes the model itself flexible and allows you to update its shape when you have new observations on the system, without having to throw away everything that was "learned" until then. In addition, it seems that the Bayesian inference allows, to a certain extent, to treat approximate models of nonlinear systems, which is deemed to be practically impossible with other approaches. [14]

In a previous path we have noticed how machine learning techniques are widely used in text analysis and interpretation tasks. We have briefly described the *Natural Language ToolKit* (*NLTK*), a didactically oriented library of algorithms written in the Python programming language: it devotes an entire section to machine learning; *spaCy*, a Python library a little more modern and more professional, is based on updated *deep learning* algorithms. And we saw that machine learning algorithms are widely applied in computational linguistics tasks such as

- *segmentation* of texts into sentences and *tokens*
- document *classification* and *sentiment analysis*
- calculation of stylistic and / or semantic *similarity* between documents
- sentence by sentence *alignment* and, within certain limits, word by word *alignment* of parallel texts
- construction of *translation memories* and automatic translation on a statistical basis (*SMT = statistical machine translation*).

BIBLIOGRAPHY AND WEBOGRAPHY

[1] Critical thinking about average. From top drawer teachers - resources for teachers of mathematics
https://topdrawer.aamt.edu.au/Statistics/Assessment/Assessment-tasks/Critical-thinking-about-average

[2] Steve Skiena, Probability versus Statistics. In Calculated Bets: Computers, Gambling, and Mathematical Modeling to Win!
https://www3.cs.stonybrook.edu/~skiena/jaialai/excerpts/node12.html

[3] A very Scottish history of insurance
http://sonsofscotland.com/scottish-history-insurance/

[4] Yuval Noah Harari, Sapiens, Da animali a dèi – Breve storia dell'umanità, Bompiani, 2018

[6] Università di Bologna – Progetto matematica, Definizione assiomatica o la teoria unificata di probabilità.
http://progettomatematica.dm.unibo.it/ProbElem/8definiz.html

[7] Nelson Education Ltd., Probability Events. 2009
http://www.nelson.com/school/elementary/mathK8/quebec/0176237879/documents/NM8SB_12B.pdf

[8] YouMath - Scuola di matematica e fisica, Eventi compatibili, incompatibili e complementari.
https://www.youmath.it/lezioni/probabilita/probabilita-discreta/1200-eventi-compatibili-e-incompatibili.html

[9] Giulio D'Agostini, Incertezze in Fisica e nelle altre scienze naturali. Istituto Nazionale di Fisica Nucleare

https://www.roma1.infn.it/~dagos/PRO/node5.html

[10] Giancarlo Ragozini, Università di Napoli Federico II, Principali teoremi del calcolo delle probabilità.
http://www.federica.unina.it/sociologia/statistica/principali-teoremi-del-calcolo-della-probabilita/

[11] J.B. Dow, Early actuarial work in eighteenth-century Scotland. Lecture delivered to the Faculty of Actuaries in Scotland, 16th October 1972.
https://www.researchgate.net/publication/305950556_Early_actuarial_work_in_eighteenth-century_Scotland

[12] Inferenza statistica
https://ccrma.stanford.edu/~apinto/Inferenza_statistica.pdf

[13] Giovanni Barazzetta, Il teorema di Bayes e la probabilità condizionata.
https://library.weschool.com/lezione/bayes-teorema-formula-esercizi-probabilita-condizionata-14965.html

[14] When machine learning meets complexity: why Bayesian deep learning is unavoidable
https://medium.com/neuralspace/when-machine-learning-meets-complexity-why-bayesian-deep-learning-is-unavoidable-55c97aa2a9cc

[15] Wikipedia, Bayesian inference.
https://en.wikipedia.org/wiki/Bayesian_inference