*STATISTICA, anno LXXVII, n. 1, 2017*

# BAYESIAN STATISTICAL INFERENCE

Bruno De Finetti (1977)

*Translation of the paper published in the Proceedings of the Congress on "I fondamenti dell'inferenza statistica", held in Florence on the* $28^{th}$*–*$30^{th}$ *April 1977, Dipartimento Statistico, Università degli Studi di Firenze, 1978.*

## 1. What remains to be said?

The major difficulty I encounter (which, I believe, more or less everybody will encounter) in deciding what to say on this topic, consists in the impression that everything has been said over and over again, from all different viewpoints and with all the possible and imaginable nuances, in conformity with, or different from, or opposite to the viewpoints of each of us.

Then what remains to be said?

As far as I am concerned, I am afraid that I have said and written even too much; anyway, as I do wish to fulfill the kind and welcome invitation by the organisers of this Congress and I believe that, because such a Congress is held, it is good for all the opinions to be heard, I have accepted and I will try, if not to say new and different things, to shed more light on the analysis of the role played by inductive reasoning (*i.e.* Bayesian) within the area of decision theory, considered as a distinct element of the chain of considerations forming, as a whole, the rational procedure to be followed in order to choose a decision in the best possible way.

Considered as a link of such a chain, the moment of statistical induction on one hand results as autonomous due to the peculiarity of its function, and on the other it results necessarily conditioned and univocally defined as a linking element between the knowledge we start from and the consequent choice of the decision proved as the most suitable. This linking element, which responds to a specific and well-defined function of its — but, in my opinion, one which cannot be ignored or substituted — consists, in fact, in the application of Bayes' theorem.

## 2. The "black box"

The opposite method (including more or less all the usual procedures of the so-called objectivistic statistics) aims on the contrary to merge all the phases into an undifferentiated, complex and undivided block, hence leading to the proliferation of those mechanical recipes which indefinitely enrich the statistical recipe book: those recipes which, according to the appropriate definition introduced by Irving J. Good, simply represent "adhockeries". That is to say they are formalistic and empirical ad hoc methods used to draw conclusions from a set of data, like rabbits from a magician's hat, or, more concretely, to make decisions only on the basis of the result of those observations which gave such data.

Yet more, the sphere of competences attributed to the "black box" is even broader, because the prescription of methods followed for the collection of data to be introduced as input is generally assigned to it as well.

In other words, adhockeries are decision methods operating as a black box, as a mysterious and unprecise object into which data and questions are introduced at one end and ready-made answers and advice are obtained from the other end. This, without any rational or intuitive justification, i.e. the "why" linking input to output of the logical mechanism. Under such conditions, even less so may we know or say if and why the suggestions of a black box are acceptable or how to choose among contrasting suggestions emerging from different black boxes.

## 3. Elements and steps of a decision process

A decision process consists in choosing, from two or more or infinite possible alternatives, the one which is considered as the most suitable. As a first approximation, let us say the one leading to the maximum gain obtained (or, in any case, the least loss, if nothing better). This is a first approximation because here we refer to the gain in monetary terms, while, more precisely, we should express ourselves in utility terms (as we will see shortly).

If, on the basis of the already acquired information and of the opinions formed from these, a person must or wishes to decide straight away, he cannot do anything but to compare the gain that each decision might procure him, and choose the one resulting in maximum gain. This indeed is only valid in the banal and extreme situation where there is no uncertainty, or if the gain corresponding to each choice were certain and known by the concerned individual, that is, the person who has to make the decision (whom we will call, as usual, the "decision maker").

Naturally, mentioning such a banal case first only has the aim of introducing later, as we proceed, the elements which make the decision problems interesting and really significant. The first important element is the uncertainty by which the probabilistic aspect is introduced, which, however, is not the one which is more strictly related to our theme: inductive reasoning, and more specifically the Bayesian one.

Having to choose among $n$ alternatives, $A_1, ..., A_i, ..., A_n$ it is obvious that, in the banal and already mentioned case when one knows the certain values of gain $S_1, ..., S_i, ..., S_n$ corresponding to each of the alternatives, he will choose the one, we call it $A_h$, for which $S_h$ is maximum ($S_h$). In the case where $S_i$ are aleatory (that is: they are unknown for sure), in order to compare the preferabilities, it would be necessary to evaluate the equivalent true value, that is, the value $\bar{S}_i$ for which the concerned person would judge indifferently the choice between the uncertain value $S_i$ and the certain $\bar{S}_i$.

This value might be the forecast of $S_i$, $P(S_i)$ (that is, according to other terminologies, the mean value or the mathematical expectation of the gain): at any rate, the certain value $\bar{S}_i$ will usually be much smaller than $P(S_i)$, depending upon the level of risk aversion of the concerned individual and the amount of money at stake in relation to his wealth.

However, all this does not influence the mechanism of the choice: the alternative to be chosen will always be the one - $A_h$ - for which $\bar{S}_h$ results the greatest amongst the $\bar{S}_i$.

(Whenever the maximum value is reached for many alternatives ($A_{h'}, A_{h''}$, *ecc.*) it would be indifferent to choose any one of them as $A_h$ or even any linear (convex) combination, that is the choice, irrelevant in such a case, of one of them could be linked to an extraction).

## 4. Measurements in utility terms

As well known, the logical criterion of such choices is explained by introducing the utility notion, as it has been done since '700 by Daniel Bernoulli and other contemporaries. It is a matter of introducing function $U(x) =$"Utility" of owning a capital $x$ (according to the individual to whom we are referring), that is, the function corresponding to the following meaning: the individual judges as acceptable a betting in both the meanings, only and only when the forecast of his utility increment (not of the gain) from $U(x)$ to $U(x+S)$ is nil: that is

$$P(U[x+S]) = P(U[x]) = U(x)$$

(In the limit case where the risk aversion was nil, one would fall into the usual formulation valid for small risks. Then $U(x) = x$ and the condition for a fair betting is the usual $P(S) = 0$, that is gain forecast nil: in fact in such a case it is

$$P(x+S) = P(x) = x, \qquad P(S) = P(x+S) - P(x) = x - x = 0$$

This demonstration is reported *ad abundantiam*, although everything is even too obvious, only to show that such a case, as a particular case, falls back into the more complex formulation of the general case.

This had to be stated beforehand so as not to forget that, in the strict sense, the reasoning should always be in utility terms; but now we observe that for us this would be a useless complication, an irrelevant one for the purpose of the conclusions which interest us and concern the inference (and particularly the Bayesian inference). Hence, leaving any discussion on utility out of consideration, that is reasoning as if the benefit of a gain was given by the very gain, or as if (always remember this understood "as if" in order to avoid misunderstandings!), as if (I was saying) we were speaking of an individual showing no risk aversion, that is one likely to simply maximise the gain forecast (in monetary terms).

It must be very clear that, in all this, induction has no bearing: it comes into play only when there is the possibility (and eventually the convenience, depending on the cost) to acquire information as to make a better decision afterwards, that is to decide on the basis on the new level of information if and which possible bettings to make because they result as advantageous in the light of it.

Every coherent procedure by itself must only be Bernoullian (in the aforesaid sense of leading to the maximum of the expected utility); it would be inappropriate to call it Bayesian because such a definition relates only to a possible accessory ingredient, however important, of the decision process. Precisely, the

Bayesian character only concerns the possible updating of probability evaluations by the addition of new, relevant information regarding the problem data. Considering the title of this paper, this is the matter which we must focus our attention on.

## 5. The value of information

So far we have not spoken of information, this not because it did not matter, but only because it was implicitly understood (as usual as long as we refer to probability evaluations by a given individual in a given time, based on the level of information he has then).

The information level, $H$, is the logical product, or intersection, of all the propositions known at that time as true (or certain) by the individual. On the basis of this level the individual will value the probability of any event $E$, so that writing $P(E)$ should rigorously be written $P(E|H)$: probability of $E$ according to the individual with the information $H$.

Perhaps it is better (but it is a mere psychological trick) to speak in the first person (probability that I assign to $E$ in my present level of information), or better still (according to the suggestion and the use by Leonard J. Savage) in the second person (probability that You assign to event $E$ in your present level of information $H$). Such a rule should make the reader feel that judgements and reasonings are his own; as a habit it might perhaps become too artificial and heavy, but I will follow it to give the definition of the value of an information.

First of all, let us define that a value of information does not exist as such: what exists is the value it has for an individual who, on its basis, may choose among such actions and decisions $D_J$, which are available to him, in a wiser way taking into account further elements of judgement.

By definition, this value will be the expected gain given by the possibility of deciding, taking into account the information under consideration, hence $p^*$ is the maximum price considered worth paying for the same information.

In order to strengthen the reasoning, let us assume that You may, by paying a specific price $p$, learn which one is the true subhypothesis $H_r$ among the (incompatible, exhaustive) $H_1, H_2, ..., H_n$ (such that $H_1 + H_2 + ... + H_n = H$). Let be $H_r$. It is clear that such information will be of advantage omly if the expected benefit obtained by it surpasses cost $p^*$.

The benefit depends on the possibility to make, in any verified hypothesis $H_r$, the decision $D_r$ which correspondingly is the most advantageous one, instead of choosing always the one which is the best, when any further information is lacking. The sum of the maximum values is greater than the maximum of the sum, unless the decomposition of the hypotheses $H_i$ is not totally irrelevant. From this the convenience of adjusting the decision to the knowledge of further circumstances which are relevant for the purposes of evaluation. But a real benefit is only obtained if the expected gain is greater than the cost $p^*$ for acquiring the information.

I want to clarify that a benefit is obtained only in the sense that the expected gain (in utility terms) results as positive. Naturally, the actual gain may always become a loss. But this is an obvious and general matter as shown by banal examples. For example, the person who, thanks to an obtained observation, would choose to bet on a white ball in the drawing from an urn, when he has been informed that the percentage of white balls is maximum, certainly has (at the time) achieved a more advantageous position, but it might well happen that just one of the very few black balls, perhaps the only one, will be drawn. Still more commonly, it is certainly of advantage (before the drawing) to exchange one lottery ticket with ten tickets, but it might well be that ultimately the opposite will be true, because that very single ticket wins the first prize while the others don't. These are banal and obvious observations, but often, when one pursues complex mathematical arguments or metaphysical lucubrations, the sense of such suggestions may well easily escape one's attention; the shrewdness of thinking in concrete and practical terms may well be stifled by the dominance and weight of high-sounding theories and of terrifying formulations.

## 6. Practical examples

The best way is always to think of interesting and practical examples, examples which do not allow for a translation into phraseologies and formalisms of a schematic and abstract type with the consequent trap of suggesting hurried and simplistic applications of mechanical and stereotyped small formulae.

If one wants to choose not a stereotyped but a practical and interesting example to illustrate the value of an information for decisions in uncertain conditions, it is better to think of situations like that of a restaurant which has to collect its supplies for the next day. The complete information would consist in knowing, exactly and with certainty, how many people will come to eat, having also previously agreed on the *menu* (as in some banquets). A rather complete information would consist in knowing only the number of people with the uncertainty (greater or smaller depending on whether it is the case of faithful customers or occasional ones or foreigners, *etc.*) regarding which type of food and in what quantity to prepare (or be able to prepare) hoping to be able to satisfactorily face all the requirements without risking excessive leftovers. The least information is that which, lacking any specific information, may be made by synthetic induction from one's own personal experience, on the basis of the usual average frequency and taking into account factors that may come into play (*e.g.* forecast of good or bad weather; festivities or attendance to performances or sport events, *etc.*, nearby; probable arrival of individual tourists or groups, and so on).

The value of each information (in relation to facts such as those quoted, *e.g.*, in the specific case) consists in decreasing the risk of losses (non-gains included) deriving from errors of approximation in forecasts, and the information value (in the introduced technical sense) is the price that would be fair for the restaurant keeper to pay a hypothetical fortune-teller who would be able to tell him for sure which dishes and in what quantity will be requested.

This would be a total, complete piece of information (here and almost everywhere a purely chimerical one); any partial information has a lower value, as understood in the same way: it represents the forecasts of saving (or, equivalently, of the maximum gain) which the restaurant keeper may achieve if he takes it into account when planning purchases personnel costs *etc.*.

One may observe that the value is always positive, or at the most nil: in the worst case, in fact, one may decide without taking the additional information into account. It may well be (obviously, and as we have already noted) that the gain following the information, after all, is also randomly negative, that is it leads to a greater loss: in fact such a risk always exists (in spite of the advantage in forecasting); the judgement of total advantage takes everything into account, this as well.

If the acquisition of the information involves a cost (as it is generally, either in the real sense of expense or as labour, thinking, *etc.*) its advantage (that is, of its acquisition) is usually given by the difference between value and cost. The optimum decision (as far as the choice of the information to ask for) is obviously that for which the difference between value and cost is maximum.

## 7. On bayesian reasoning

The previous observations were made as a preface to our discussion of Bayesian reasoning, thus allocating it at the right place, that is, explaining if and where and how and to which extent it is useful.

Personally, anyway, I would say (or I would like to say) not "Bayesian reasoning" but "the so-called Bayesian reasoning", at the cost of decreasing (in a certain way) the *status* attributed to it, mainly by the opponents.

My dislike of such a definition — of "Bayesian reasoning" or "Bayesian induction" — does not certainly mean disagreement or insufficient adhesion to Bayes' position, but, on the contrary, I believe it is the only correct one, and by adding a superfluous adjective one may arise a doubt that other, non-erroneous, acceptable forms of inductive reasoning exist.

I could repeat, as mine, a sentence by J. Cornfield (in his "Presidential address" at the American Statistical Society, 1974, full of critical suggestions and practical observations): "Bayes' theorem is important because it provides an explication for this process of consistent choice between hypotheses on the basis of observations and for quantitative characterization of their respective uncertainties".... "Actually, Bayes' result follows so directly from the formal definitions of probability and related concepts that it is perhaps overly solemn to call it a theorem at all". I have developed the same concept (much more widely and quoting Cornfield and others) in the "Invited review paper" at the ISI Congress in Vienna, 1973, by the title of *Bayesianism: Its unifying role for both the foundations and applications of statistics.*

In other words, I have expressed this position of mine, mentioning it at the last lesson (29.11.76) on the occasion of my retirement. I oppose the expression "Bayesian induction", not because of an insufficient adhesion to Bayes' position, but, on the contrary, because according to me it is the only correct one, and the addition of a superfluous adjective may raise doubts that there are other non-erroneous, acceptable forms of inductive reasoning. He who says "Bayesian induction", for coherence should also call, according to me, "Pythagorean arithmetics" the one which, to carry out a product calculation, accepts respecting the traditional Pythagorean table, admitting that 6 times 8 is 48, or 3 times 9 is 27, *etc.* (while others, according to some other different fashion, might prefer that 6 times 8 is 90 or 3 times 9 is 77).

Nothing different intervenes in Bayes' theorem and the most intuitive and natural illustration can be given by resorting to the very well-known (and perhaps even misused) diagrams (or "potatoes") system by Eulero-Venn.

Let us think of any drawn outlines ("potatoes" to use the more expressive term), interlinked or disconnected, internal to a square (which represents the certain event and the area of which is taken as unit). The area delimited by each of them represents one of the events — *e.g.* $E_1$, $E_2$, $E_3, ...,$ $E_6$ — and the parts common to two or more of them represent their logical products. Let us assume that we made certain that the area of each piece is equal to the related probability (square area – certain event – taken egual to 1)[1].

If we now consider a further event $H$ and we represent it in the same way, probabilities $P(H \cap E_i)$, $P(H \cap E_i E_j)$, *etc.* will be represented by the part of $E_i$, of $E_i, E_j$ *etc.* within the new "total potato" $H$; and if we assume to come to know that $H$ is true (and nothing else, otherwise we will land with an $H^{'}$ which is contained in $H$), and we ask ourselves which is probability $P(E|H)$ of an event $E$ (*e.g.* $E_i$, $E_i, E_j$ *etc.*) subordinately to the hypothesis of $H$ occurring, it is given in the same way, but considering only the areas within potato $H$ and setting them equal to its area (which becomes the unity when $H$ is assumed as certain).

Thus $P(E|H) = P(EH)/P(H)$ (conditional probability theorem) and also, symmetrically, $P(H|E) = P(EH)/P(E)$,

$$P(E|H) = P(E)\frac{P(H|E)}{P(H)} \qquad \text{(Bayes' theorem)}$$

As one can see, the so-called Bayes' theorem simply states that nothing changes when the information change: all what happens is that the probability of the events, which end up excluded, become nil and the probabilities of the surviving events modify themselves, as what remains is only the portion of area included in $H$. However as a compensation they relate to the only area of $H$.

At a glance, however, the formula says this in an immediately intuitive way, without difficulties and without words.

## 8. Conclusion

The conclusion may appear disappointing for those who love abstruse and pretentious things and disdain simple, obvious, intuitive ones, but I feel it is difficult, even impossible, to justify such an attitude either towards the theory and conceptual vision or in the practice of applications.

Is it diminishing to think, like Bertrand, that probability theory is nothing else but common sense reduced to calculation? Would it make sense, would it be justifiable to judge as more precious any artificial cavil and trust it against the evidence of considerations in conformity to common sense?

Whoever loves the far-fetched "adhockeries" may find that a precise and so-simple method like the Bayesian one has to be deplored because it "makes Statistics dull" (as has been written by Herman Chernoff, a clever statistician, I think one having a Bayesian tendency, but a bit too attached to the availability of a rich collection of alternative recipes considered to be more or less classic).

On the other hand, is it perhaps diminishing towards Bayes to say that his formula is obvious? Certainly not. Formulae are worth what they are worth,

[1] One could also leave this out of consideration assuming that, in each part, weights or collectives are allocated which are proportional to probabilities and the total collective is taken as unit.

but the essential matter is to guess, to foresee and, in the end, to completely explain all the meaning they bear. Great is Bayes' merit: he was the first to see the right road, with justified hesitations; the same is true for Laplace who, *vice versa*, overcame such hesitations in a far too simple way; jumping to the present times, Abraham Wald's great merit must be pointed out because he, although not a Bayesian, did characterise through the notion of admissibility those procedures which formally respond to the Bayesian formulation.

In fact, Wald's result says, in simple words, that any method of statistical decision is either Bayesian (even though not intentionally) or it may be substituted by a Bayesian method which is unconditionally preferable to it. Using an image, one could say that the Bayesian methods are the surface points, limits of what may be reached on the basis of data and related hypotheses. Superior methods do not exist; other methods remain below the surface. To these methods, those that are within the cone formed starting from them towards the surface are preferable, and among them those (Bayesian) on the surface itself are the best (they cannot be further improved on the basis of available data and statements).

In fact, Wald's result says, in simple words, that any method of statistical decision is either Bayesian (even though not intentionally) or it may be substituted by a Bayesian method which is unconditionally preferable to it. Using an image, one could say that the Bayesian methods are the surface points, limits of what may be reached on the basis of data and related hypotheses. Superior methods do not exist; other methods remain below the surface. To these methods, those that are within the cone formed starting from them towards the surface are preferable, and among them those (Bayesian) on the surface itself are the best (they cannot be further improved on the basis of available data and statements).

Such being the situation, it happens that (as observed by Lindley: in Making Decisions, p. 43) anyone who accepts the admissibility notion (as defined by the non-Bayesian Abraham Wald) but refuses the use of initial probability (often improperly called a priori) puts himself in an impossible situation: if he uses an admissible procedure, it is the case of a Bayesian type rule corresponding (perhaps without his awareness) to a special choice of initial probability; otherwise this does not happen but his choice is not admissible. Lindley thus concludes: "although these results do not show how such initial probabilities must be assigned, they demonstrate that responsible behaviour is equivalent to their determination and vice versa. This explains the revival of interest for Bayes and it makes one think that it is more than a temporary fancy" (p. 44). "To the non-frequentist (Cornfield says, p. 46), the use of such like inadmissible procedures seems to be an extravagant price to pay for the support of a philosophical position, and an empirically unverifiable one at that".

I add a recent contribution to such a debate, in a letter by T. Leonard (Univ. of Warwick) to "News and Notes" of the Royal Statistical Society (Nov. 1976). He notes that no conclusion can be drawn with no method (either explicitly Bayesian or pretendingly objective) from a statistical analysis of a finite number of data: this would be meaningless, senseless, without any initial information included. "If we all admit that such an initial information must be used, why not adopt, or at least approximate, a formal mechanism for its inclusion?".

As for the objection that evaluations such as those of initial probabilities, because subjective, vary from individual to individual, one can note two things. First, they often differ very little among themselves and from a certain *communis opinio*, suggested by various circumstances, and second, convergence increases, as one obtains more information. Lindley says (p. 32) “there is nothing in our argument that makes agreement inevitable, but in practice it will often happen that agreement can be reached, given enough evidence”. This, in the more specific case, is the same as Bayesian reasoning, but, evidently, here it hints at the same phenomenon in a broader sense of refinement of the intuitive vision which is useful as a guide.

The notion of exchangeability, which is my main contribution, is also inspired to this purpose for indicative information see the above mentioned last lesson of mine: it is not a mathematical artifice (as it seems when it is called “de Finetti’s representation theorem”), but the instrument used “to present induction as a very natural way of reasoning on probabilities of observable facts, avoiding metaphysical pseudoentities and obscurities” (*Bayesianism, p. 361*).

## References

J. Cornfield (1967). Review Int. St. Inst., 35, 34–49.

B. De Finetti (1973). Bull. of the Int. Institute, Proc. of the $39^{th}$ sess. of the ISI, vol. 4.

D.CV. Lindley (1971). *Making Decisions.* Wiley Intersc., London.

A. Wald (1939). Annals Math. Stat., 10, 299–326.

## Summary

This work was translated into English and published in the volume: Bruno De Finetti, Induction and Probability, Biblioteca di Statistica, eds. P. Monari, D. Cocchi, Clueb, Bologna, 1993. Bayesian statistical Inference is one of the last fundamental philosophical papers in which we can find the essential De Finetti’s approach to the statistical inference.

*Keywords*: Bayesian inference; initial probabilties; statistical decision