Introduction to Bayesian Statistics

Bayesian Statistics and subjective probability, Prior Distribution, Current Approaches (part 3)

Teacher: Matilde Trevisani

DEAMS

A.A. 2024/2025

(aggiornato: 2025-03-18)

Agenda (about 3 lectures)

Introduction to Bayesian Statistics

- Context
- Machinery of Probability
- Interpretations of Probability
- Direct and Inverse problems
- Bayes Theorem
- Modern (classical) Statistics
- Bayesian Statistics and subjective probability
- Prior Distributions
- Current Approaches
- Final Notes

Bayesian Statistics and subjective probability (continue)

XX century

Bayesian statistics was rediscovered in the XXth century.

because of new questions

- breaking the *Enigma* code during WW2 (see *Edward Simpson: Bayes at Bletchly Park*, http://mathcenter.oxford.emory.edu/site/math117/bayesTheorem/)
- combining historical and current information in setting insurance rates (the actuarial technique known as credibility theory turns out to be based on Bayesian reasoning)
- estimating the probability of events such as
  - probability of an aviation accident involving two planes (in the 50s)
  - probability of an accidental explosion of an H-bomb

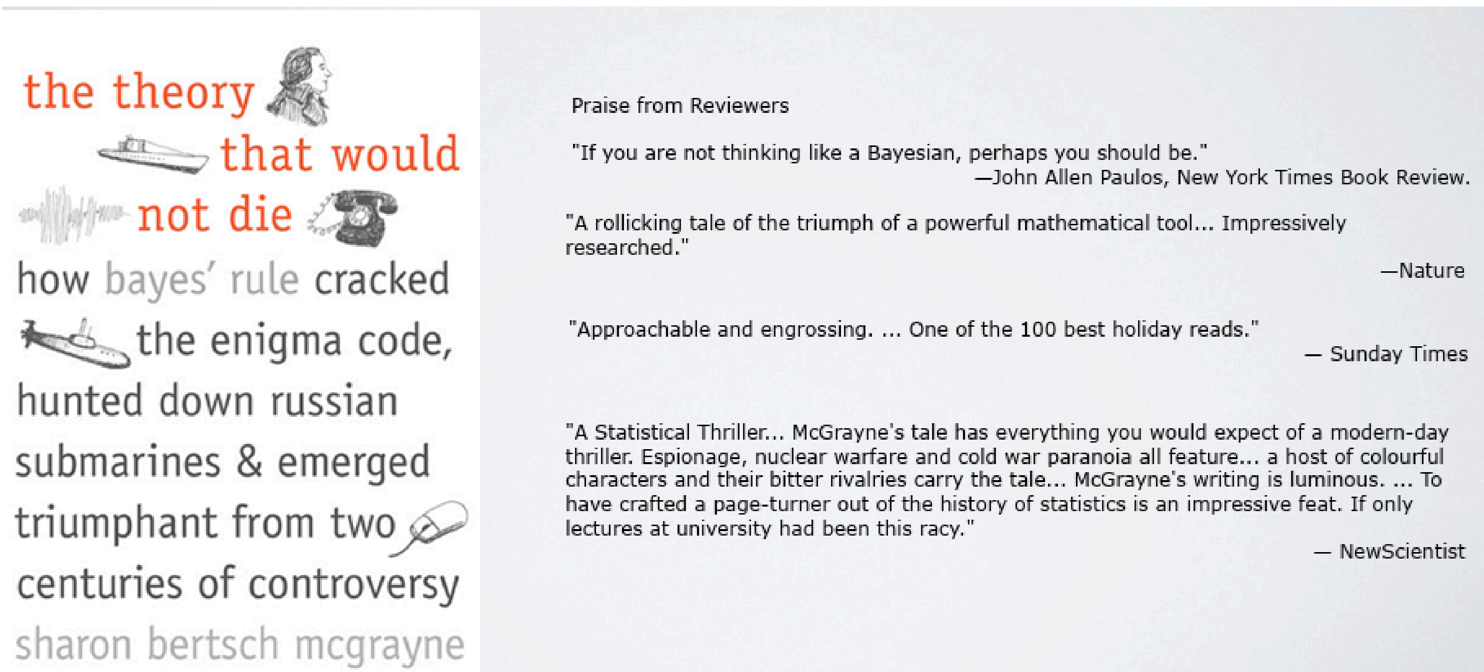
thanks to the availability of computers

- Bayesian analytical results are available only for simple problems
- Bayesian computation is rather intensive (Monte Carlo methods are fundamental)

Bayes' Rule is just conditional probability
So why is it such a big deal?

McGrayne (2011)

> A vivid account of the generations-long dispute over Bayes' rule, one of the greatest breakthroughs in the history of applied mathematics and statistics



Different questions, different methods

Behind the choice of the preferred statistical approach, frequentist or Bayesian, there might be the question which is asked and the information available.

Fisher (1890-1932), working in genetics, was actually performing experiments

- no need for a prior information
- easy to frame the interpretation in the repeated sampling paradigm

Many applications of Bayesian inference in this period involved the need/desire to

- **combine different sources of information**
- **assess probability of events which were never observed**

Problems with probability

Although employed in special contexts in XX century, Bayesian Statistics did not achieve general acceptance (far from it).

One of the reasons why Bayesian statistics was difficult to accept is related to the frequentist definition of probability.

It is conceptually difficult to frame a distribution on the model (parameter) as a frequentist probability.

We need another probability!

Let us take a step back and discuss about this.

Limits of the frequentist interpretation of probability

Frequentist definition of probability applies to a narrow set of events, those which can be embedded, at least ideally, in a sequence of repetitions.

This is easily done for situations such as the toss of a coin with a head and a tail.

However we can easily think of "events" for which it does not work:
(each one for a different reason)

- Win the lottery in the next draw
- The Italian women's volleyball team wins the next World Championship (not strictly applicable)
- WWIII between the USA and Russia will break out next year (***one-off*** event)
- The coin shows heads when we know the coin is either double-headed or double-tailed, but we don't know which (a hypothesis is either true or false)
- The number of non-European citizens in Italy on 01/01/2025

The last two cases neither describe an event (a possible outcome of a random experiment).

The model (and any *gear* of it) is a random thing

In statistical applications, the situation is analogous to the two-headed/two-tailed coin example.

Suppose you want to estimate *the number* N *of non UE citizens in Italy on 1/1/2025*, N is then a random number in the Bayesian setting, one may object that

- N is not a random quantity (it is intrinsically a fixed number, albeit unknown);
- I cannot specify a probability distribution on a non random quantity!

The frequentist definition does not help in interpreting a probability distribution on N.

What is probability?

We can free us from the frequentist interpretation by taking the axiomatic definition of probability.

Probability (Definition)

Probability is a measure on a set of events (outcomes) such that

- it is non negative
- it is additive over mutually exclusive events
- sums to 1 over all possible mutually exclusive outcomes

This does tell us nothing about how it is measured and what can be used for.

We will all agree that we can use it to describe limiting relative frequencies of occurrence of events in repeated sequences, we may not agree on whether we can use it for something else.

Do we need another probability?

First, should we use it for something else?

We may take the stance that only events for which a sequence of ideal repetition is thinkable are permitted.

This is unsatisfying intuitively and practically since we have to deal with more general kinds of uncertainty (and they are relevant, think the H-bomb accidents) and we do routinely deal with them, that is we do take decisions based on some evaluation of such uncertain (non repeatable) events (think betting or weather forecasts).

Still, we might say that this kind of events is dealt with by **common sense** and is out of scope for a formal treatment by probability, but it might also be the case that probability could describe how common sense works.

Common sense: deductive logic $\to$ plausible logic

An example of common sense is an inference like

$$\left.\begin{array}{c}\text{if } A \text{ then } B \\ A \text{ is true}\end{array}\right\} \Rightarrow B \text{ is true}$$

which is described by **deductive logic**.

We also do inferences like the following

$$\left.\begin{array}{c}\text{if } A \text{ then } B \\ B \text{ is true}\end{array}\right\} \Rightarrow A \text{ more plausible}, \qquad \left.\begin{array}{c}\text{if } A \text{ then } B \\ A \text{ is false}\end{array}\right\} \Rightarrow B \text{ less plausible}$$

or even

$$\left.\begin{array}{c}\text{if } A \text{ then } B \text{ is more plausible} \\ B \text{ is true}\end{array}\right\} \Rightarrow A \text{ is more plausible}$$

This is a common type of reasoning (**plausible logic**), even in everyday life.

It is sensible to try to describe it, that is, to quantify less/more plausible.

Probability as extension of true-false logic

If the aim is to represent the state of uncertainty on a fact'', then conditional probability is the only system which satisfies the following axioms

I. States of uncertainty are represented by real numbers.

II. Qualitative correspondence with common sense.

- If the truth value of a proposition increases, its probability must also increase.
- In the limit, small changes in propositions must yield small changes in probabilities.

III. Consistency with true-false logic.

- Probabilities that depend on multiple propositions cannot depend on the order in which they are presented.
- All known propositions must be used in reasoning -- nothing can be arbitrarily ignored.
- If, in two settings, the propositions known to be true are identical, the probabilities must be as well.

Coherence of bets (remind)

Another "proof" that probability as defined by the axioms is the only reasonable way to describe uncertainty is the Dutch book argument.

Let us define the probability of an event P(E) as:

- The price you would pay in exchange for a return of 1 if the event occurs, and 0 otherwise.
- The price you would accept in exchange for having to pay 1 if the event occurs, and 0 otherwise.

In other words, once you state P(E) you would buy or sell the random amount |E| in exchange for P(E).

Suppose that you assess probabilities for a set of events, then if your probabilities do not satisfy the axioms it is possible to devise a combination of bets leading to a sure loss (gain). (That is, there is a combination of bets such that you would loose money no matter what happens.)

Probability to describe uncertainty

To some, these considerations make using probability to represent uncertainty a compelling choice and so Bayesian reasoning (which is a consequence of probability) the only reasonable way to update information (probabilities).

> Bayesian Statistics offers a rationalist theory of personalistic beliefs in contexts of uncertainty, with the central aim of characterising how an individual should act in order to avoid certain kinds of undesirable behavioural inconsistencies (Bernardo and Smith).

This way of looking at probability leads quite naturally to the subjective definition of probability.

Subjective probability

> Probabilities are states of mind and not states of nature.
>
> -- *Leonard J. Savage, 1954*

One then accepts that the probability is not an objective property of a phenomenon but rather the opinion of a person and one defines

Definition (subjective probability)

> The probability of an event is, for an individual, his degree of belief on the event.

Bruno de Finetti (c. 1906-1985), Italian probabilist and actuary (for Generali) in *Theory of probability* (1970) he wrote

> Probability does not exist

and proposes the subjective definition of probability and the coherence framework, based on the bet interpretation.

If the probability is a subjective degree of belief, it depends on the information which is subjectively available, and it is also clear that **by random we mean not known for lack of information**.

Then, a parameter is random because is unknown.
(In frequentist approach a "true" parameter, generally unknown, exists and one tries to “estimate” it.)

- Frequentists define probability as the long-run frequency of a certain measurement or observation. The frequentist says that there is a single truth and our measurement samples noisy instances of this truth. The more data we collect, the better we can pinpoint the truth. The archetypal example is the repeated tossing of a coin to see what the long-run frequency is of heads or tails; this long-run frequency then asymptotes to the truth.

- Bayesians define probability as the plausibility of a **hypothesis** given incomplete knowledge. To a Bayesian there is no Platonic truth out there which we want to access through data collection (or perhaps we should say there may be a Platonic truth, but it will always remain outside our experience). For a Bayesian, there is just data which we can use as evidence for particular hypotheses. A Bayesian coin-tosser just observes a series of coin tosses and then uses this information to make deductions about, for example, how likely it is that the coin is fair.

The real-world coin-toss experiment lives in the Bayesian domain: the experiment is done under limited knowledge of the initial conditions and precise set-up. Given that limited knowledge, how do we use the experimental outcomes to evaluate a statement such as “the coin is fair”? This is a Bayesian question. Bayesian statistics is the statistics of the real world, not of its Platonic ideal.

Bayesian statistics and subjective probability

The subjective definition of probability is most compatible with the Bayesian paradigm, stated as follows

- the parameter to be estimated is a well specified quantity but is not known for lack of information
- a probability distribution is (subjectively) specified for the parameter to be estimated, this is called the *prior distribution*
- after seeing experimental results the probability distribution on the parameter is updated using Bayes' theorem to combine experimental results (likelihood) and prior distribution to obtain the *posterior distribution*.

Subjective probability and Bayesian update rule (which is actually a consequence of probability rules) establish a system to describe inference whose input are the prior beliefs and the data and the output is updated (posterior) beliefs.

Uses of probability

- To describe variation
    - using aleatory or phenomenological or descriptive probability
    - analogous to saying probability of rolling a 3 with an apparently fair die is 1/6.
- To quantify uncertain knowledge
    - by making epistemic or inferential statements
    - analogous to saying you are 90% sure that Greenland is part of the Kingdom of Denmark.

Bayesian inference

- Founded on an inferential principle of equivalence: there is only **one kind of probability** for both inferential and descriptive purposes, with inferential statements being obtained from descriptive ones merely by applying the laws of probability.

Subjective Bayes

This approach is sometimes called **subjective Bayes**, it had a lot of followers since the 60s (see Lindley (1970), Lindley (2013), Savage (1972)).

Some probably took it for something more than it is (could be: when people are captivated by its spell, they tend to proselytize, and exaggerate its virtues).

In fact, for many it became the only coherent foundation of statistics, whereas the alternatives (Fisher, Neymann-Pearson and alike) looked like a collection of *ad hoc* tools lacking a proper justification.

This lead to the formation of two factions each rejecting the methods of the other, on part of the anti Bayesians the criticism were focused on the fact that admitting a subjective nature of the conclusions made them useless from a scientific point of view.

Even if we accept that subjective Bayes is a good description of reasoning under uncertainty in broad sense, it is still relevant to discuss whether this is acceptable in a scientific context: simplifying a bit, the role of prior distributions is central to this.

Prior distribution

Need for prior

A critical issue in Bayesian inference (and one of the reasons why it did not get acceptance at the beginning) is the need for prior information.

While classical statistics is only concerned with the information coming from the data, Bayesian statistics is a rule to update information based on the data: we must start somewhere.

This was seen as a major issue since it introduces an element of subjectivity in the analysis.

This will be discussed later, we make now two preliminary notes concerning

- where the prior comes from;
- the subjectivity (in the sense of arbitrariness) of results.

Source of prior information

Think of the female birth example again, but with the following sample:

> In 2010 in Muggia (small city near Triest) 38 males and 47 females were born.

According to likelihood inference (for θ, pr. of a female birth)

- the ML estimate is $\hat{\theta} = 0.553$
- the 95% c.i. is [0.441, 0.659]
- the p-value for $H0 : \theta \geq 0.5$ is 0.8

What do you think of this information?

You probably think something along the lines of "This sample tells me nothing".

Why is that? Well, because you have, in fact, prior information.

We usually have prior information

In fact, it would be rare that we model a situation were we have no prior information at all.

Prior information may come from

- substantive knowledge about the process generating the data (we may be unsure about the exact mechanism but we usually know something),
- observations made in the past.

With this in mind, the prior distribution should not look so strange.

Subjectivity of results: vanishing priors

People do not like prior information because they do not like that two persons with the same data may reach different conclusions because they start from different prior informations.

While this is true, it is also true that, if the prior information is not unreasonable, the conclusions tend (asymptotically) to become equal as more data are gathered.

We will discuss what does "unreasonable" means, but the basic requirement is that we do not exclude any possibility (by assigning it a null prior probability).

Moreover, we will discuss how to distinguish prior distributions with respect to how much they weigh on the conclusion (how informative they are): there are methods to ensure that the conclusions are less influenced by the prior.

Subjectivity of results: standard priors

What we have said above assumes that the prior can (and should) represent the beliefs prior to the observations.

It is also possible to take a different approach, within the Bayesian paradigm.

In the example of female birth Laplace assumed a uniform prior on θ: he viewed this as a way to express indifference with respect to the possibilities.

This is kind of reasonable, although problematic for some aspects, the idea can be made more precise.

Subjectivity of results: make the prior irrelevant

The idea is that the prior does not need to convey information, rather it is regarded as a technical component of the model.

This idea lies behind the so called

- non informative priors
- reference priors

whose name tells it all, although maybe too optimistically:

- "informativeness" is not a well defined concept, beware of attaching a precise meaning to the intuitive idea
- the posterior still depends on the prior

With these caveats let's say that particular distributions can be defined to avoid the subjective interpretation of the prior distribution.

This approach is sometimes called **objective Bayes** (or automatic Bayes).

Recap

- Probability calculus
  direct problems ↔ P(Data | Model)

- First statistical inference is "Bayesian" (Bayes, Laplace)
  probability of causes ↔ P(Model | Data)

- B. approach put aside (for philosophical and technical reasons)
  classical statistics (Gosset, Fisher and many others)

  - likelihood
  - repeated sampling

- interest on BS renewed (new problems, computational advances)
  There are issues related to the subjectivity of the results

  - subjective B. is put forward as a compelling paradigm (de Finetti, Lindley, Savage)
  - objective B. to compromise between classical and B. stat (Jeffreys)

Current approaches

Present day

- We still lack a clear foundation of statistical inference which is agreed upon.
- This is not only an abstract issue, it has been argued that it is at the root of practical problems in applications of statistics: the issue of hypotheses testing in applied science (see Nuzzo (2014); Goodman (2016), see also Pauli (2018) for an overview of the issue).
- In what follows we will see some modern overviews of the scenario on the foundations of statistics (Senn, Efron and Hastie, Royall).

Map of approaches by Senn

Senn (2011) maps the various approaches we have briefly considered according to whether they focus on

- direct or inverse probabilities on one hand;
- inference or decision on the other hand (here the latter means that we are interested in the consequences of using a certain criterion).

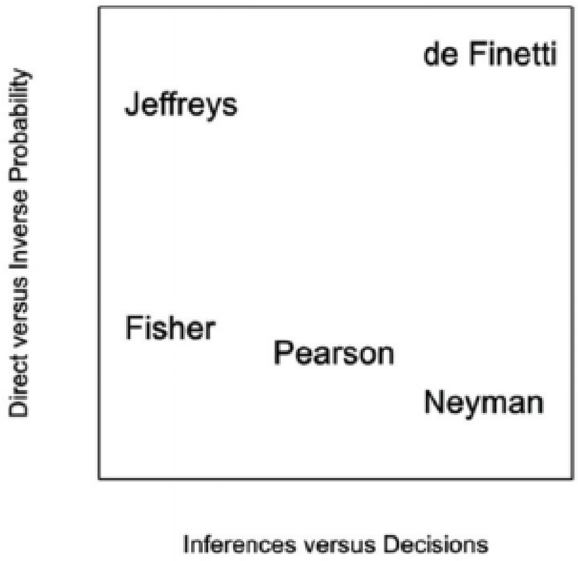
(Keep in mind that any scheme like this is bound to oversimplify.)

Classic Statistics

- Likelihood: Fisher
- Hypothesis Test: Neyman-Pearson

Bayesian Statistics

- Objective: Jeffrey
- Subjective: de Finetti



Senn Map

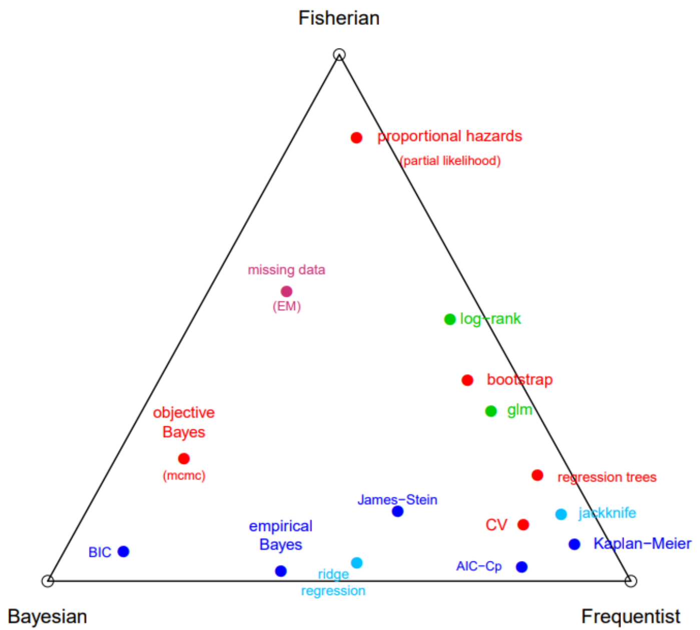
Map of techniques/approaches by Efron and Hastie

Efron and Hastie (2021) map some statistical techniques with respect to their guiding principles and the **relevance of the computational aspect**.

The triangle is not meant to give a complete picture: it shows a **selection of techniques**, as described in the text (*Early Computer-Age Methods*: 15 major topics, **1950s through 1990s**), and the Bayesian, frequentist, and Fisherian influences on them. It shows how these techniques are **based on a mixture of approaches**

Color indicates the **importance of electronic computation** in their development ->

It would not be easy to place some XXI century machine learning developments (a philosophically atheistic approach to statistical inference) here (of course they would all be red)



- red, crucial;
- violet, very important;
- green, important;
- light blue, less important;
- blue, negligible.

Some quotes from *Computer age statistical inference, student edition: algorithms, evidence, and data science*

> Statistical inference is an unusually wide-ranging discipline, located as it is at the triple-point of mathematics, empirical science, and philosophy. The discipline can be said to date from 1763, with the publication of Bayes' rule. The most recent quarter of this 250-year history—from the 1950s to the present—is the **"computer age"**, the time when computation, the traditional bottleneck of statistical applications, became faster …. The book is an examination of how statistics has evolved over the past sixty years.
>
> The role of electronic computation is central to our story. This doesn't mean that every advance was computer-related. [But] Almost all topics in twenty-first-century statistics are now computer-dependent.
>
> Very broadly speaking, algorithms are what statisticians do while inference says why they do them.
>
> A particularly energetic brand of the statistical enterprise has flourished in the new century, **data science**, emphasizing algorithmic thinking rather than its inferential justification.

Royall: different approaches for different questions

Royall (2004) distinguishes methods based on the question they seek to answer.

Three questions can be asked to the data

1. What should I believe?
2. How should I behave?
3. What is the evidence?

Royall stance is that

(3) is answered by the likelihood alone,

(1) is answered by the posterior (needs the likelihood and the prior),

(2) needs the posterior and the costs of errors.

Note that (2) is different from the "decision" Senn has in mind.

Mixing

The good note is that the factions are no more: to some extent at least, statisticians are willing on taking what is relevant from each approach.

> It seems quite clear that both Bayesian and frequentist methodology are here to stay, and that we should not expect either to disappear in the future. ... Philosophical unification of the Bayesian and frequentist positions is not likely, nor desirable, since each illuminates a different aspect of statistical inference.
> -- *Bayarri and Berger (2004)*

- In practice this has meant that many Bayesians now consider it reasonable to assess model adequacy (which is incoherent with viewing (posterior distribution on) models as beliefs, as beliefs cannot be wrong)
- Frequentist properties of Bayesian procedures are studied. See hybridization

> An assessment of strengths and weaknesses of the frequentist and Bayes systems of inference suggests that ... inferences under a particular mode! should be Bayesian, but model assessment can and should involve frequentist ideas (Little, 2006)

Today

- On pragmatic grounds, it is reasonable to use whatever approach is best suited for the situation at hand, this is the most common attitude among applied statisticians.

> "Pure" Bayes, "pure" frequentist, "pure" any statistical philosophy, pairs nicely with Port, but when you leave port for the high seas of applications, some degree of impurity is usually necessary. Consequently, statisticians who engage in important studies use their paradigm as an aid to navigation, not as a straightjacket. The goal is to do a good job, and one can't be (too) doctrinaire
>
> -- *Tom Louis, 2019*

- It is also reasonable to interpret Bayesian techniques as modelling techniques rather than a philosophical stance (thus disconnecting it from the subjective interpretation), in this sense the role of the prior can be downplayed, from a source of information to a regularization device (a part of a model). See Object Bayes

- Whatever attitude you will take, keep in mind, however, that the Bayesian approach is the only correct one (on a probabilistic reasoning ground) and all other procedures are justified only as approximations of the Bayesian ones.

Bayesian reasoning beyond statistics

Compelling nature of Bayesian reasoning

Recall that

> Bayesian Statistics offers a rationalist theory of personalistic beliefs in contexts of uncertainty, with the central aim of characterising how an individual should act in order to avoid certain kinds of undesirable behavioural inconsistencies (Bernardo and Smith).

we noted that this has lead some to argue for taking Bayesian reasoning as the foundation of statistical inference.

In fact, we have said that Bayesian reasoning (as a probabilistic reasoning) could be the paradigm to extend deductive logic to plausible logic.

Role of Bayesian reasoning

These circumstances lead some to think that Bayesian reasoning could (should) be used as the paradigm of inductive logic, that is, beyond its statistical scope: a recipe for human reasoning in general.

Let us then consider contexts were beliefs are important (central) and discuss to what extent Bayesian reasoning fits practice:

- diagnostic: where interest lies on whether a tested person is ill
- law: where interest lies in the belief on guilt or innocence of a defendant.
- science (epistemology): where interest lies in the truth of a theory

The question is whether (to what extent, under which conditions) Bayesian reasoning can describe (model) the reasoning process of a scientist (judge/juror, clinician) who accept/rejects theories (decides over guilt/innocence, diagnose patients).

Diagnostic example: fully appropriate

Law example: Here the reasoning works, the point is that most if not all the probabilities which are involved have to be elicited and are debatable.

Scienze : works when there are clear alternative theories.

How well does B reasoning describe reasoning?

The tenet is that Bayesian reasoning would work if we could precisely define all alternative theories a priori, which is unrealistic in general: *"because it is very hard to be sufficiently imaginative and because life is short."*.

Bayesian reasoning can not describe all human reasoning.

Bayesian reasoning is a compelling framework but only

- limited to the hypothesis under consideration (and limited by the reasonableness of such hypotheses),
- conditional on the likelihood given to the evidence under the various hypothesis.

In statistical terms this translates in conditional on the model specification, hence the importance of evaluating the fit to check on model adequacy.

Further readings

For the **history of Bayesian statistics**, with examples, see McGrayne (2011).

For a modern presentation of the **subjective Bayes** approach see Jaynes (2003) and Lindley (2013), further readings include De Finetti (1974), Jeffreys (1998), Lindley (1970) and Savage (1972).

The **classical approach** to inference is described in Cox (2006), its principles are discussed in Mayo and Cox (2006) and Mayo (2011). The works which originated the approach are also readable although with some difficulty: Fisher (1922), Fisher (1925) and Neyman, Pearson, and Pearson (1933).

A **modern approach to Bayesian inference** is in Gelman, Stern, Carlin et al. (2013), see also Gelman and others (2011) and Gelman and Shalizi (2013) for the role of Bayesian inference in science.

References

Bayarri, M. J. and J. O. Berger (2004). "The interplay of Bayesian and frequentist analysis".

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge university press.

De Finetti, B. (1974). *Theory of probability*. John Wiley & Sons.

Efron, B. and T. Hastie (2021). *Computer age statistical inference, student edition: algorithms, evidence, and data science*. Vol. 6. Cambridge University Press.

Fisher, R. (1922). "On the mathematical foundations of theoretical statistics". In: *Phil. Trans. R. Soc. Lond. A* 222.594-604, pp. 309-368.

Fisher, R. (1925). *Statistical methods for research workers.* Oliver and Boyd, Edinburgh.

Gelman, A. and others (2011). "Induction and deduction in Bayesian data analysis". In: *Rationality, Markets and Morals* 2.67-78, p. 1999.

Gelman, A. and C. R. Shalizi (2013). "Philosophy and the practice of Bayesian statistics". In: *British Journal of Mathematical and Statistical Psychology* 66.1, pp. 8-38.

Gelman, A., H. S. Stern, J. B. Carlin, et al. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Goodman, S. N. (2016). "Aligning statistical and scientific reasoning". In: *Science* 352 (6290), pp. 1180-1181.

References (continue)

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Lindley, D. V. (1970). *Introduction to probability and statistics from a Bayesian viewpoint*.

Lindley, D. V. (2013). *Understanding uncertainty*. John Wiley & Sons.

Mayo, D. G. (2011). "Statistical science and philosophy of science: Where do/should they meet in 2011 (and beyond)?"

Mayo, D. G. and D. R. Cox (2006). "Frequentist statistics as a theory of inductive inference". In: *Optimality*. Institute of Mathematical Statistics, pp. 77-97.

McGrayne, S. B. (2011). *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press.

Neyman, J., E. S. Pearson, and K. Pearson (1933). "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706, pp. 289-337. DOI: 10.1098/rsta.1933.0009.

Nuzzo, R. (2014). "Scientific method: Statistical errors". In: *Nature* 506.7487, pp. 150-152. ISSN: 0028-0836.

Pauli, F. (2018). "The p-value Case, a Review of the Debate: Issues and Plausible Remedies". In: *Studies in Theoretical and Applied Statistics*. Ed. by C. Perna, M. Pratesi and A. Ruiz-Gazen. Cham: Springer International Publishing, pp. 95-104. ISBN: 978-3-319-73906-9.

References (continue)

Royall, R. (2004). "The likelihood paradigm for statistical evidence". In: *The nature of scientific evidence: Statistical, philosophical, and empirical considerations*, pp. 119-152.

Savage, L. J. (1972). *The foundations of statistics*. Dover Publications.

Senn, S. (2011). "You may believe you are a Bayesian but you are probably wrong". In: *Rationality, Markets and Morals* 2.48-66, p. 27.

Final Notes

The Art of Probabilistic Modeling

The art of probabilistic modeling is describing in a mathematical form (model and priors) what we already know and what we don't know.

The "easy" part is using Bayes' rule to update uncertainties

- computational challenges

Other parts of the art of probabilistic modeling are, for example,

- Model checking: does the data conflict with our prior knowledge?
- presentation: presenting the model and results to the application experts

Advantages of the Bayesian approach

- Integrate uncertainties to focus on the parts of interest
- Use relevant a priori information
- Hierarchical models
- Check and evaluation of models

Computation

We need to be able to compute the expected value with respect to the posterior distribution $p(\theta|y)$

$E_{\theta|y} [g(\theta)] = \int p(\theta|y)g(\theta)d\theta$

- Analytical
  - only for very simple models
- Computational approximations with a finite number of function evaluations
  - *grid, importance sampling*, Monte Carlo, Markov chain Monte Carlo
  - generic
- Distributional approximations
  - e.g. Laplace, variational, Expected value propagation
  - less generic, but can be much faster with sufficient accuracy

Marginal Notes

Frequentist Evaluations in Bayesian Statistics

Bayesian theory considers both epistemic and aleatory probabilities.

"Frequentist" evaluations focus on "frequentist" properties given the model and the random repetition of an observation.

- Asymptotic Consistency
- Correctness:
  - Less crucial in Bayesian inference (as small errors are more important)
- Efficiency:
  - A small quadratic error is desirable.
  - Other utility/cost functions may also be considered.
- Calibration
  - A posterior interval at α% contains the true value in α% of cases.
  - A predictive interval at α% contains future true values in α% of cases.
  - *Approximate* calibration with shorter intervals for likely true values is more important than exact calibration with longer intervals for all possible values.

See Mixing

Frequentist Statistics

Frequentist statistics accepts only aleatory probabilities

- Estimates are based on the data
- The uncertainty of the estimates is based on all possible datasets that could have been generated by the data generation mechanism

Estimates are derived to satisfy frequentist properties

- Maximum likelihood satisfies only asymptotic frequentist properties
- Most common desired properties are 1) correctness, 2) minimum variance, 3) calibration of the confidence interval
- The requirement of correctness can lead to increased variance or nonsense estimates (e.g., a correct estimate for a strictly positive parameter can be negative)
- The confidence interval is said to contain the true value in α% of the cases of repeated generation of data by the data generation mechanism
    - it does not say how likely it is that the true value is within the interval constructed on the observed data
    - it is not necessary that it is useful for having a perfect calibration

Frequentist vs Bayes vs others

There are a lot of frequentist techniques that are very useful

- for simple models and with a lot of data there is not much difference

Bayesian inference

- simpler for complex models, e.g. hierarchical
- simpler when the model changes
- a consistent way to add a priori information

Most machine learning techniques are neither purely frequentist nor purely Bayesian