1

Bayesian Inference

Before discussing Bayesian inference, we recall the fundamental problem of statistics: "The fundamental problem towards which the study of Statistics is addressed is that of inference. Some data are observed and we wish to make statements, inferences, about one or more unknown features of the physical system which gave rise to these data" (O'Hagan, 2010). Upon more careful consideration of the foundations of statistics we find many different schools of thought. Even leaving aside those that are collectively known as classical statistics, this leaves several choices: objective and subjective Bayes, fiducialist inference, likelihood based methods, and more.[1]

This diversity is not unexpected! Deriving the desired inference on parameters and models from the data is a problem of induction, which is one of the most controversial problems in philosophy. Each school of thought follows its own principles and methods to lead to statistical inference. Berger (1984) describes this as: "Statistics needs a: 'foundation', by which I mean a framework of analysis within which any statistical investigation can theoretically be planned, performed, and meaningfully evaluated. The words 'any' and 'theoretically' are key, in that the framework should apply to any situation but may only theoretically be implementable. Practical difficulties or time limitations may prevent complete (or even partial) utilisation of such framework, but the direction in which 'truth' could be found would at least be known". The foundations of Bayesian inference are better understood when seen in contrast to those of its mainstream competitor, classical inference.

[1] Subjective Bayes is essentially the subject of this volume. In addition to these schools of thought, there are even half-Bayesians who accept the use of a priori information but believe that probability calculus is inadequate to combine prior information with data, which should instead be replaced by a notion of causal inference.

1.1 The Classical Paradigm

Classical statistics seeks to make inference about a population starting from a sample. Let x (or $x = (x_1, x_2, \ldots, x_n)$, where n is a sample size,) denote the data. The set $\mathcal{X}$ of possible samples x is known as the sample space, usually $\mathcal{X} \subseteq \mathbb{R}^n$. Underlying classical inference is the recognition of variability across samples, keeping in mind that the observed data are only one of many – possibly infinitely many – data sets that could have been observed. The interpretation of the data depends not only on the observed data, but also on the assumptions put forward about the process generating the observable data. As a consequence, the data are treated as a realization of a random variable or a random vector X with a distribution F_θ, which of course is not entirely known. However, there is usually some knowledge (theoretical considerations, experimental evidence, etc.) about the nature of the chance experiment under consideration that allow one to conjecture that F_θ is a member of a family of distributions $\mathcal{F}$. This family of distributions becomes the statistical model for X. The assumption of a model is also known as the *model specification* and is an essential part of developing the desired inference.

Assuming that X is a continuous random variable or random vector, it is common practice to represent the distributions $\mathcal{F}$ by their respective density functions. When the density functions are indexed by a parameter θ in a parameter space Θ, the model can be written as $\mathcal{F} = \{f(x \mid \theta),\ x \in \mathcal{X} : \theta \in \Theta\}$. In many cases, the n variables $(X_1, X_2, \ldots, X_n)$ are assumed independent conditional on θ and the statistical model can be written in terms of the marginal densities of X_i, $i = 1, 2, \ldots, n$:

$$\mathcal{F} = \{f(x \mid \theta) = \Pi_{i=1}^n f_i(x_i \mid \theta) : \theta \in \Theta\}, x \in \mathcal{X},$$

and $f_i(\cdot \mid \theta) = f(\cdot \mid \theta)$, $i = 1, 2, \ldots, n$, if additionally the variables X_i are assumed to be identically distributed. The latter is often referred to as random sampling.

Beyond the task of modeling and parametrization, classical inference includes many methods to extract conclusions about the characteristics of the model that best represents the population and tries to answer questions like the following: (1) Are the data x compatible with a family $\mathcal{F}$? (2) Assuming that the specification is correct and that the data are generated from a model in the family $\mathcal{F}$, what conclusions can be drawn about the parameter θ_0 that indexes the distribution F_θ that "appropriately" describes the phenomenon under study?

Classical methods – also known as frequentist methods – are evaluated

under the principle of repeated sampling, that is, with respect to the performance under infinitely many hypothetical repetitions of the experiment carried out under identical conditions. One of the aspects of this principle is the use of frequencies as a measure of uncertainties, that is, a frequentist interpretation of probability. See , Paulino et al. (2018, section 1.2), for a review of this and other interpretations of probability.

In the case of parametric inference, in answer to question (2) above, we need to consider first the question of point estimation, which, *grosso modo*, is: Given a sample $X = (X_1, X_2, \ldots, X_n)$, how should one "guess," estimate, or approximate the true value θ, through an estimator $T(X_1, X_2, \ldots, X_n)$. The estimator should have the desired properties such as unbiasedness, consistency, sufficiency, efficiency, etc.

For example, with $\mathcal{X} \equiv \mathbb{R}^n$, the estimator $T(X_1, X_2, \ldots, X_n)$ based on a random sample is said to be centered or unbiased if

$$E\{T \mid \theta\} = \int_{\mathbb{R}^n} T(x_1, x_2, \ldots, x_n) \Pi_{i=1}^n f(x_i \mid \theta)\, dx_1 dx_2 \ldots dx_n = \theta, \ \forall \theta \in \Theta.$$

This is a property related to the principle of repeated sampling, as can be seen by the fact that it includes integration over the sample space (in this case $\mathbb{R}^n$). Considering this entire space is only relevant if one imagines infinitely many repetitions of the sampling process or observations of the n random variables $(X_1, X_2, \ldots, X_n)$. The same applies when one considers other criteria for evaluation of estimators within the classical paradigm. In other words, implicit in the principle of repeated sampling is a consideration of what might happen in the entire sample space.

Parametric inference often takes the form of confidence intervals. Instead of proposing a single value for θ, one indicates an interval whose endpoints are a function of the sample,

$$(T^*(X_1, X_2, \ldots, X_n), T^{**}(X_1, X_2, \ldots, X_n)),$$

and which covers the true parameter value with a certain probability, preferably a high probability (typically referred to as the confidence level),

$$P\{T^*(X_1, X_2, \ldots, X_n) < \theta < T^{**}(X_1, X_2, \ldots, X_n) \mid \theta\} = 1 - \alpha,$$

$0 < \alpha < 1$. This expression pre-experimentally translates a probability of covering the unknown value θ to a random interval (T^*, T^{**}) whose lower and upper limits are functions of $(X_1, X_2, \ldots, X_n)$ and, therefore, random variables. However, once a specific sample is observed (i.e., post-experimentally) as n real values, $(x_1, x_2, \ldots, x_n)$, this becomes a specific

interval on the real line (now with real numbers as lower and upper limits).

$$(T^*(x_1, x_2, \ldots, x_n), T^{**}(x_1, x_2, \ldots, x_n)),$$

and the probability

$$P\{^*T(x_1, x_2, \ldots, x_n) < \theta < T^{**}(x_1, x_2, \ldots, x_n) \mid \theta\} = 1 - \alpha,$$

$0 < \alpha < 1$, is no longer meaningful. In fact, once θ has an unknown, but fixed, value, this probability can only be 1 or 0, depending upon whether the true value of θ is or is not in the real interval

$$(T^*(x_1, x_2, \ldots, x_n), T^{**}(x_1, x_2, \ldots, x_n)).$$

Of course, since θ is unknown, the investigator does not know which situation applies. However, a classical statistician accepts the frequentist interpretation of probability and invokes the principle of repeated sampling in the following way: If one imagines a repetition of the sampling and inference process (each sample with n observations) a large number of times, then in $(1 - \alpha)$ 100% of the repetitions the numerical interval will include the value of θ.

Another instance of classical statistical inference is a parametric hypothesis test. In the course of scientific investigation one frequently encounters, in the context of a certain theory, the concept of a hypothesis about the value of one (or multiple) parameter(s), for example in the symbols

$$H_0 : \theta = \theta_0.$$

This raises the following fundamental question: Do the data $(x_1, x_2, \ldots, x_n)$ support or not support the proposed hypothesis? This hypothesis is traditionally referred to as the null hypothesis. Also here the classical solution is again based on the principle of repeated sampling if one follows the Neyman–Pearson theory. It aims to find a rejection region W (critical region) defined as a subset of the sample space, $W \subset \mathcal{X}$, such that

$$(X_1, X_2, \ldots, X_n) \in W \;\Rightarrow\; \text{rejection of } H_0,$$

$$(X_1, X_2, \ldots, X_n) \notin W \;\Rightarrow\; \text{fail to reject } H_0.$$

The approach aims to control the probability of a type-I error,

$$P\{(X_1, X_2, \ldots, X_n) \in W \mid H_0 \text{ is true}\},$$

and minimize the probability of a type-II error,

$$P\{(X_1, X_2, \ldots, X_n) \notin W \mid H_0 \text{ is false}\}.$$

What does it mean that the critical region is associated with a type-I error, equal to, for example, 0.05? The investigator can not know whether a false or true hypothesis is being rejected when a particular observation falls into the critical region and the hypothesis is thus rejected. However, being a classical statistician the investigator is convinced that under a large number of repetitions and if the hypothesis were true, then only in 5% of the cases would the observation fall into the rejection region. What does it mean that the critical region is associated with a type-II error equal to, say 0.10? Similarly, when a particular observation is not in the rejection region and thus the hypothesis is not rejected, then the investigator cannot know whether a true or false hypothesis is being accepted. Being a classical statistician, the investigator can affirm that under a large number of repetitions of the entire process and if the hypothesis were in fact false, only in 10% of the cases would the observation not fall into the rejection region.

In the following discussion, it is assumed that the reader is familiar with at least the most elementary aspects of how classical inference approaches estimation and hypothesis testing, which is therefore not discussed here in further detail.

1.2 The Bayesian Paradigm

For Lindley, the substitution of the classical paradigm by the Bayesian paradigm represents a true scientific revolution in the sense of Kuhn (1962) The initial seed for the Bayesian approach to inference problems was planted by Richard Price when, in 1763, he posthumously published the work of Rev. Thomas Bayes titled *An Essay towards Solving a Problem in the Doctrine of Chances*. An interpretation of probability as a degree of belief – fundamental in the Bayesian philosophy – has a long history, including J. Bernoulli, in 1713, with his work *Ars Conjectandi*. One of the first authors to define probabilities as a degree of beliefs in the truth of a given proposition was De Morgan, in *Formal Logic*, in 1847, who stated: (1) probability is identified as a degree of belief; (2) the degrees of belief can be measured; and (3) these degrees of belief can be identified with a certain set of judgments. The idea of coherence of a system of degrees of belief seems to be due to Ramsey, for whom the behavior of an individual when betting on the truth of a given proposition is associated with the degree of belief that the individual attaches to it. If an individual states odds or possibilities (chances) – in favor of the truth or untruth – as $r : s$, then the degree of belief in the proposition is, for this individual, $r/(r+s)$. For Ramsey, no set of bets in given propositions is admissible for a coherent individual if it would

lead to certain loss. The strongest exponent of the concept of personal probabilities is, however, de Finetti. In discussing the Bayesian paradigm and its application to statistics, one must also cite Harold Jeffreys, who, reacting to the predominantly classical position in the middle of the century, besides inviting disapproval, managed to resurrect Bayesianism, giving it a logical basis and putting forward solutions to statistical inference problems in his time. From there the number of Bayesians grew rapidly and it becomes impossible to mention all but the most influential – perhaps Good, Savage, and Lindley.

The well-known Bayes' theorem is a proposition about conditional probabilities. It is simply probability calculus and is thus not subject to any doubts. Only the application to statistical inference problems is subject to some controversy. It obviously plays a central role in Bayesian inference, which is fundamentally different from classical inference. In the classical model, the parameter $\theta, \theta \in \Theta$, is an unknown but fixed quantity, i.e., it is a particular value that indexes the sampling model or family of distributions $\mathcal{F}$ that "appropriately" describes the process or physical system that generates the data. In the Bayesian model, the parameter θ, $\theta \in \Theta$, is treated as an unobservable random variable. In the Bayesian view, any unknown quantity – in this case, the parameter θ – is uncertain and all uncertainties are described in terms of a probability model. Related to this view, Bayesians would argue that initial information or *a priori* information – prior or external to the particular experiment, but too important to be ignored – must be translated into a probability model for θ, say $h(\theta)$, and referred to as the prior distribution. The elicitation and interpretation of prior distributions are some of the most controversial aspects of Bayesian theory.

The family $\mathcal{F}$ is also part of the Bayesian model; that is, the sampling model is a common part of the classical and the Bayesian paradigms, except that in the latter the elements $f(x \mid \theta)$ of $\mathcal{F}$ are in general assumed to also have a subjective interpretation, similar to $h(\theta)$.

The discussion of prior distributions illustrates some aspects of the disagreement between Bayesian and classical statisticians. For the earlier, Berger, for example, the subjective choice of the family $\mathcal{F}$ is often considered a more drastic use of prior information than the use of prior distributions. And some would add: In the process of modeling, a classical statistician uses prior information, albeit in a very informal manner. Such informal use of prior information is seen critically under a Bayesian paradigm, which would require that initial or prior information of an investigator needs to be formally stated as a probability distribution on the random variable θ. Classical statisticians, for example, Lehmann, see an

important difference between the modeling of $\mathcal{F}$ and the specification of $h(\theta)$. In the earlier case one has a data set $x = (x_1, x_2, \ldots, x_n)$ that is generated by a member of $\mathcal{F}$ and can be used to test the assumed distribution.

To understand the Bayesian point of view, recall that for a classical statistician *all* problems that involve a binomial random variable X can be reduced to a Bernoulli model with an unknown parameter θ that represents a "success" probability. For Bayesians, each problem is *unique* and has its own real context where θ is an important quantity about which there is, in general, some level of knowledge that might vary from problem to problem and investigator to investigator. Thus, the probability model that captures this variability is based on *a priori* information and is specific to a given problem and a given investigator. In fact, *a priori* information includes personal judgements and experiences of most diverse types, resulting from in general not replicable situations, and can thus only be formalized in subjective terms. This formalism requires that the investigator comply with coherence or consistency conditions that permit the use of probability calculus. However, different investigators can in general use different prior distributions for the same parameter without violating coherence conditions.

Assume that we observe $X = x$ and are given some $f(x \mid \theta) \in \mathcal{F}$ and a prior distribution $h(\theta)$. Then Bayes' theorem implies[2]

$$h(\theta \mid x) = \frac{f(x \mid \theta)h(\theta)}{\int_\theta f(x \mid \theta)h(\theta)\, d\theta}, \ \theta \in \Theta, \tag{1.1}$$

where $h(\theta \mid x)$ is the posterior distribution of θ after observing $X = x$. Here, the initial information of the investigator is characterized by $h(\theta)$, and modified with the observed data by being updated to $h(\theta \mid x)$. The denominator in (1.1), denoted $f(x)$, is the marginal (or prior predictive) distribution for X; that is, for an observation of X whatever the value of θ.

The concept of a likelihood function appears in the context of classical inference, and is not less important in the Bayesian context. Regarding its definition, it is convenient to distinguish between the discrete and continuous cases (Kempthorn and Folks, 1971), but both cases lead to the function of θ,

$$\begin{aligned} L(\theta \mid x) &= kf(x \mid \theta), \quad \theta \in \Theta \quad \text{or} \\ L(\theta \mid x_1, \ldots, x_n) &= k\Pi_i f(x_i \mid \theta), \quad \theta \in \Theta, \end{aligned} \tag{1.2}$$

which expresses for every $\theta \in \Theta$ its likelihood or plausibility when $X = x$ or $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ is observed. The symbol k represents a

[2] Easily adapted if x were a vector or if the parameter space were discrete.

factor that does not depend on θ. The likelihood function – it is not a probability, and therefore, for example, it is not meaningful to add likelihoods – plays an important role in Bayes' theorem as it is the factor through which the data, x, updates prior knowledge about θ; that is, the likelihood can be interpreted as quantifying the information about θ that is provided by the data x.

In summary, for a Bayesian the posterior distribution contains, by way of Bayes' theorem, all available information about a parameter:

prior information + information from the sample.

It follows that all Bayesian inference is based on $h(\theta \mid x)$ [or $h(\theta \mid x_1, x_2, \ldots, x_n)$].

When θ is a parameter vector, that is, $\theta = (\gamma, \phi) \in \Gamma \times \Phi$, it can be the case that the desired inference is restricted to a subvector of θ, say γ. In this case, in contrast to the classical paradigm, the elimination of the nuisance parameter ϕ under the Bayesian paradigm follows always the same principle, namely through the marginalization of the joint posterior distribution,

$$h(\gamma \mid x) = \int_{\Phi} h(\gamma, \phi \mid x) d\phi = \int_{\Phi} h(\gamma \mid \phi, x) h(\phi \mid x) d\phi. \tag{1.3}$$

Possible difficulties in the analytic evaluation of the marginal disappear when γ and ϕ are *a priori* independent and the likelihood function factors into $L(\theta \mid x) = L_1(\gamma \mid x) \times L_2(\phi \mid x)$, leading to $h(\gamma \mid x) \propto h(\gamma) L_1(\gamma \mid x)$.

1.3 Bayesian Inference

In the Bayesian approach, it is convenient to distinguish between two objectives: (1) inference about unknown parameters θ, and (2) inference about future data (prediction).

1.3.1 Parametric Inference

In the case of inference on parameters, we find a certain agreement – at least superficially – between classical and Bayesian objectives, although in the implementation the two approaches differ. On one side, classical inference is based on probabilities associated with different samples, x, that could be observed under some fixed but unknown value of θ. That is, inference is based on sampling distributions that "weigh" probabilistically the values that a variable X or statistic $T(X)$ can assume across the sample

space. On the other hand, Bayesian inference is based on subjective probabilities or *a posteriori* credibilities associated with different values of the parameter θ and conditional on the particular observed x value. The main point is that x is fixed and known and θ is uncertain.

For example, once x is observed, a Bayesian being asked about the hypothesis $\{\theta \leq 0.5\}$ would directly address the question by calculating $P(\theta \leq 0.5 \mid x)$ based on $h(\theta \mid x)$, i.e., without leaving probability calculus. In contrast, a classical statistician would not directly answer the question. Stating, for example, that the hypothesis $H_0 : \theta \leq 0.5$ is rejected at significance level 5% does not mean that its probability is less than 0.05, but that if the hypothesis H_0 were true, (i.e., if in fact $\theta \leq 0.5$), then the probability of X falling into a given rejection region W would be $P(X \in W \mid \theta \leq 0.5) < 0.05$, and if in fact $x \in W$, then the hypothesis is rejected.

In O'Hagan's words (O'Hagan, 2010), while a Bayesian can state probabilities about the parameters, which are considered random variables, this is not possible for a classical statistician, who uses probabilities on data and not on parameters and needs to restate such probabilities such that they seem to say something about the parameter. The question is also related to a different view of the sample space. For a classical statistician, the concept of the sample space is fundamental, as repeated sampling would explore the entire space. A Bayesian would start by objecting to the reliance on repeated sampling and would assert that only the actually observed value x is of interest and not the space that x belongs to, which could be totally arbitrary, and which contains, besides x, observations that could have been observed, but were not.[3]

In estimation problems a classical statistician has several alternatives for functions of the data – estimators – whose sampling properties are investigated under different perspectives (consistency, unbiasedness, etc.). For a Bayesian there is only *one* estimator, which specifically is the posterior distribution $h(\theta \mid x)$. One can, of course, summarize this distribution in different ways, using mode, mean, median, or variance. But this is unrelated to the problem facing a classical statistician, who has to find a so-called *optimal estimator*. For a Bayesian such a problem only exists in the context of decision theory, an area in which the Bayesian view has a clear advantage over the classical view. Related to this, Savage claims that in past decades

[3] The irrelevance of the sample space also leads to the same issue about stopping rules, something which Mayo and Kruse (2001) note, recalling Armitage, could cause problems for Bayesians.

the central problem in the face of uncertainty is shifting from *which inference one should report*, to *which decision should be taken*. As individual decisions have been considered outdated by some philosophers, we have also recently seen a resurgence of the Bayesian approach in the context of group decisions.

Under a Bayesian approach, confidence intervals are replaced by credible intervals (or regions). Given x, and once a posterior distribution is determined, one finds a credible interval for a parameter θ (assume, for the moment, a scalar). The interval is formed by two values in θ, say $[\underline{\theta}(x), \bar{\theta}(x)]$, or simpler, $(\underline{\theta}, \bar{\theta})$, such that

$$P(\underline{\theta} < \theta < \bar{\theta} \mid x) = \int_{\underline{\theta}}^{\bar{\theta}} h(\theta \mid x)\, d\theta = 1 - \alpha, \tag{1.4}$$

where $1 - \alpha$ (usually 0.90, 0.95, or 0.99) is the desired level of credibility. If $\Theta = (-\infty, +\infty)$, then one straightforward way of constructing a (in this case, central) credible interval is based on tails of the posterior distribution such that

$$\int_{-\infty}^{\underline{\theta}} h(\theta \mid x)\, d\theta = \int_{\bar{\theta}}^{+\infty} h(\theta \mid x)\, d\theta = \frac{\alpha}{2}. \tag{1.5}$$

Equation (1.4) has an awkward implication: The interval $(\underline{\theta}, \bar{\theta})$ is not unique. It could even happen that the values θ in the reported interval have less credibility than values θ outside the same interval. Therefore, to proceed with the choice of an interval that satisfies (1.4) and at the same time is of minimum size, Bayesians prefer to work with HPD (*highest posterior density*) credible sets $A = \{\theta : h(\theta \mid x_1, x_2, \ldots, x_n) \geq k(\alpha)\}$, where $k(\alpha)$ is the largest real number such that $P(A) \geq 1 - \alpha$. For a unimodal posterior, the set becomes a HPD credible interval.

Credible sets have a direct interpretation in terms of probability. The same is not true for confidence intervals, which are based on a probability not related to θ, but rather a probability related to the data; more specifically, they are random intervals based on a generic sample, and which after observing a particular sample become a *confidence* of covering the unknown value θ by the resulting numerical interval. In general, this can not be interpreted as a probability or credibility about θ. Besides other critical aspects of the theory of confidence intervals (or regions), there are the ironical comments of Lindley (1990), who says to know various axioms of probability – for example, those due to Savage, de Finetti, or Kolmogorov – but no axiomatic definition of *confidence*.

For example, when a Bayesian investigates a composite hypothesis H_0 :