An Integrated Mathematical Statistics Primer: Objective Bayesian Construction, Frequentist Evaluation

José M. Bernardo
Universitat de València, Departamento de Estadística e I.O.
Facultad de Matemáticas, Dr. Moliner 50
46100-Burjassot, Valencia, Spain
jose.m.bernardo@uv.es

Bayesian Statistics is typically taught, if at all, separately from conventional frequentist methods. It is becoming clear, however, that the emergence of powerful objective Bayesian methods (where the result, as in frequentist statistics, only depends on the assumed model and the observed data) provides a new unifying perspective on most established methods, and may be used in situations (e.g. hierarchical structures) where frequentist methods cannot. On the other hand, frequentist procedures provide mechanisms to evaluate and calibrate *any* statistical method. Hence, it may be the right time to consider an integrated approach to mathematical statistics, where objective Bayesian methods provide the inferential construction elements, and frequentist methods the necessary evaluations. The emphasis of this presentation will be on undergraduate courses on mathematical statistics, but the main ideas may also be applied to more basic introductory and service courses.

1. Introduction

A comparative analysis of the undergraduate teaching of statistics through the world shows a clear imbalance between what it is taught and what it is later needed; in particular, most primers in statistics are exclusively frequentist and, since this is often their only course in statistics, many students never get a chance to learn important Bayesian concepts which would have improved their professional skills. Moreover, too many syllabuses still repeat what was already taught by mid last century, boldly ignoring the many problems and limitations of the frequentist paradigm later discovered.

Hard core statistical journals carry today a sizeable proportion of Bayesian papers, but this does not yet translates into comparable changes in the teaching habits at universities. History often shows important delays in the introduction of new scientific paradigms into basic university teaching, but this inertia factor is not sufficient to explain the slow progress observed in the introduction of Bayesian methods into mainstream statistical teaching. When the debate flares up, those who prefer to maintain the present situation invoke mainly two arguments: (i) Bayesian statistics is described as subjective, and thus inappropriate for scientific research, and (ii) students must learn the dominant frequentist paradigm, and it is not possible to integrate both paradigms into a coherent, understandable course. The first argument only shows lack of information from those who voice it: *objective* Bayesian methods are well known since the 60's, with landmark books by Jeffreys, Lindley, Zellner, Press and Box & Tiao, and *reference analysis*, whose development started in late 70's (see *e.g.*, Bernardo Smith, 1994, Ch. 5, and references therein), provides a general methodology which includes and generalizes the pioneering solutions. The second argument is however much stronger: any professional who makes use of statistics needs to know frequentist methods (not just because of their present prevalence, but because they may be used to analyse the expected behaviour of *any* statistical methodology), and it is not easy to integrate into a single course the basic concepts of two paradigms which are often described as mutually incompatible. The purpose of this presentation is to suggest an integrated approach, where objective Bayesian methods are used to derive a unified, consistent set of solutions to the problems of statistical inference which occur in scientific

investigation, and frequentist methods (designed to analyse the behaviour under sampling of *any* statistical procedure) are used to establish the behaviour under repeated sampling of the proposed objective Bayesian methods.

Section 2 briefly describes a possible syllabus to develop these ideas in practice. Section 3 contains a simple, illustrative example.

2. An integrated approach to theoretical statistics

The central idea of our proposal is to use objective Bayesian methods to derive statistical procedures which directly address the problems of inference commonly found in scientific investigation, and to use frequentist techniques to *evaluate* the behaviour of those procedures under repeated sampling. For instance, to quote one of the simplest examples, if data consists of a random sample of size n from a normal $\mathrm{N}(x \,|\, \mu, \sigma)$, with mean $\overline{x}$ and variance s^2, the interval $\overline{x} \pm t_{\alpha/2}\, s/\sqrt{n-1}$ is obtained from an objective Bayesian perspective as a *credible region* to which (given the data) the population mean μ belongs with (rational) probability $1-\alpha$. In our experience, this type of result—which describes what may said about the quantity of interest given available information—is precisely the type of result in which scientists are genuinely interested. Moreover, the frequentist analysis of that region estimator shows that, under repeated sampling, regions of this form would contain the true value of μ in $100(1-\alpha)\%$ of the possible samples, thus providing a valuable calibration of the Bayesian result. The correspondence between the objective credible regions and the frequentist confidence regions (which is exact in this example) is nearly always approximately valid for sufficiently large samples.

What follows is the particular implementation of an integrated approach to theoretical statistics which been used for the last two years in teaching the course *Mathematical Statistics* to third year undergraduate students of mathematics at the University of Valencia, Spain.

1. *Foundations*
   - Introduction to decision theory
   - Probability as a rational conditional measure of uncertainty
   - Divergence and information measures
2. *Probability models*
   - Exchangeability and representation theorems
   - Likelihood function. Properties and approximations
   - Sufficiency and the exponential family
3. *Inference: Objective Bayesian methods*
   - The learning process. Asymptotic results
   - Elementary reference analysis
   - Point estimation as a decision problem
   - Region estimation: lowest posterior loss regions
   - Hypothesis testing as a decision problem
4. *Evaluation: Frequentist methods*
   - Expected behaviour of statistical procedures under repeated sampling
   - Risk associated to point estimators
   - Expected coverage of region estimators
   - Error probabilities of hypothesis testing procedures

It is argued that an integrated approach to theoretical statistics requires concepts from decision theory. Thus, the first part of the proposed course includes basic Bayesian decision theory, with special attention granted to the concept of probability as a rational measure of uncertainty. Divergence measures between distributions are also discussed in this module, with emphasis in the *intrinsic discrepancy*, $\delta\{p_1, p_2\} = \min[k\{p_1 \,|\, p_2\}, k\{p_2 \,|\, p_1\}]$, where $k\{p_j \,|\, p_i\} = \int_{\boldsymbol{\mathcal{X}}} p_i(\boldsymbol{x}) \log[p_i(\boldsymbol{x})/p_j(\boldsymbol{x})]\, d\boldsymbol{x}$ is the Kullback-Leibler divergence of p_j from p_i, and with the discrepancy between two families defined as the minimum discrepancy between their elements.

The second part of the course is devoted to probability models. The concept of exchangeability, and the intuitive content of representation theorems, are both described to provide students with an important mathematical link between repeated sampling and Bayesian analysis. The definition and properties of the likelihood function, the concept of sufficiency, and a description the exponential family of distributions complete this module.

The third part of the proposed syllabus is a brief course on modern objective Bayesian methods. The Bayesian paradigm is presented as a mathematical formulation of the learning process, and includes an analysis of the asymptotic behaviour of posterior distributions. *Reference priors* are presented as *consensus* priors designed to be always dominated by the data, and procedures are given to derive the reference priors associated to regular models. Point estimation, region estimation and hypothesis testing are all presented as procedures to derive useful summaries of the posterior distributions, and implemented as specific decision problems. The *intrinsic loss function*, based on the intrinsic discrepancy between distributions, is suggested for conventional use in scientific communication: the intrinsic loss $\delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, the loss from using a model in the family $\mathcal{M}_0 = \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}), \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ as a proxy for model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$, is defined as the intrinsic discrepancy $\delta\{p_{\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}},\ \mathcal{M}_0\}$ between the distribution $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ and the family of distributions $\mathcal{M}_0$. This loss function is invariant under one-to-one reparametrizations, and hence produces a unified set of solutions to point estimation, region estimation and hypothesis testing problems which is consistent under reparametrization (Bernardo, 2004), a rather obvious requirement, which unfortunately many statistical methods fail to satisfy.

The last module of the course presents the frequentist paradigm as a set of methods designed to analyse the behaviour under repeated sampling of possible solutions to problems of statistical inference. In particular, they are used to study the risk associated to point estimators, the expected coverage of region estimators, and the error probabilities associated to hypothesis testing procedures, with special attention to the behaviour under sampling of the objective Bayesian procedures discussed in the third module. The required evaluations are made using both analytical techniques, when the relevant sampling distributions are easily derived, and Monte Carlo simulation techniques when they are not.

3. An Example

To illustrate the ideas proposed, we conclude with a a simple example. Let $\overline{x}$ be the mean of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from $p(x \,|\, \theta) = \theta e^{-x\theta}$. The reference prior here is $\pi(\theta) = \theta^{-1}$, and the corresponding posterior is gamma $\pi(\theta \,|\, \boldsymbol{x}) = \mathrm{Ga}(\theta \,|\, n, n\overline{x})$, a function of the sufficient statistic $\overline{x}$. The reference posterior $\pi(\theta \,|\, \boldsymbol{x})$ is represented in the right panel of Figure 1 for a random sample of size $n = 10$, simulated with $\theta = 2$, which yielded $\overline{x} = 0.608$. The intrinsic loss of using model $p(\boldsymbol{x} \,|\, \theta_0)$ as a proxy for $p(\boldsymbol{x} \,|\, \theta)$ (whose value is independent of the parametrization chosen to describe the model) is $\delta\{\theta_0, \theta \,|\, n\} = n\, \delta_1\{\theta_0, \theta\}$, with

$$\delta_1\{\theta_0, \theta\} = \begin{cases} (\theta/\theta_0) - 1 - \log(\theta/\theta_0), & \text{if } \theta \le \theta_0 \\ (\theta_0/\theta) - 1 - \log(\theta_0/\theta), & \text{if } \theta > \theta_0. \end{cases}$$

The reference posterior expectation of $\delta\{\theta_0, \theta \,|\, n\}$ (which uses the observed data $\boldsymbol{x}$ to estimate the minimum log-likelihood ratio $\log[p(\boldsymbol{x} \,|\, \theta)/p(\boldsymbol{x} \,|\, \theta_0)]$ against $\theta = \theta_0$ to be expected under repeated sampling), is the *intrinsic statistic*

$$d(\theta_0 \,|\, \boldsymbol{x}) = \int_0^\infty \delta\{\theta_0, \theta \,|\, n\}\, \pi(\theta \,|\, \boldsymbol{x})\, d\theta \approx \tfrac{1}{2}\Big[1 + n\, \delta\{\theta_0,\, \tilde{\theta}(\boldsymbol{x})\}\Big],$$

where $\tilde{\theta}(\boldsymbol{x}) = [1 - (2n)^{-1}]\, \overline{x}^{-1}$, represented in the left panel of Figure 1. In a hypothesis testing situation, the *intrinsic k-rejection region* R_k consists of those θ_0 values such that $d(\theta_0 \,|\, \boldsymbol{x}) > k$, on the grounds that, given $\boldsymbol{x}$, the expected log-likelihood ratio against them would be larger

than k. For instance (see Figure 1), with the data described, the values of θ_0 smaller than 0.513 or larger than 4.768 yield an intrinsic statistic value larger than $k = \log(100) \approx 4.6$, and would therefore be rejected using this conventional threshold, since the expected log-likelihood ratio against them is expected to be larger than $\log(100)$. Moreover, the posterior expected loss of using an estimate $\tilde{\theta}$ as a proxy for the true value of θ is $d(\tilde{\theta} \,|\, \boldsymbol{x})$; this is minimized at the *intrinsic estimate*, $\theta^*(\boldsymbol{x}) \approx [1 - (2n)^{-1}]\, \overline{x}^{-1}$, which is the intrinsic loss Bayes estimate ($\theta^* = 1.568$ in this case, represented in both panels with a big dot).

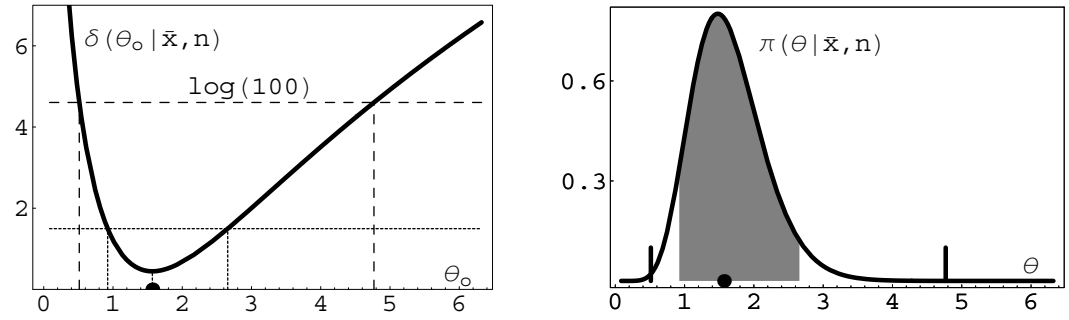

Figure 1. *Intrinsic objective Bayesian inference for an exponential parameter θ.*

A *lowest posterior loss* (LDL) p-credible region is one of the form $C_p \equiv \{\tilde{\theta}; d(\tilde{\theta} \,|\, \boldsymbol{x}) \leq k(p)\}$ and such that $\int_{C(p)} \pi(\theta \,|\, \boldsymbol{x})\, d\theta = p$. For instance, $C_{0.95}$ here consists of those parameter values with expected loss below 1.496, the interval $C_{0.95} = [0.923, 2.657]$ shaded in Figure 1. Moreover, the sampling distribution of $\overline{x}$ is gamma $p(\overline{x} \,|\, \theta, n) = \mathrm{Ga}(\overline{x} \,|\, n, n\theta)$, and the *sampling* distribution of $t(\boldsymbol{x}) = \overline{x}\,\theta$ is $p(t \,|\, \theta, n) = \mathrm{Ga}(t \,|\, n, n)$; but this is *also* the *posterior* distribution of $\phi(\theta) = \overline{x}\,\theta$, $\pi(\phi \,|\, \overline{x}, n) = \mathrm{Ga}(\phi \,|\, n, n)$. Hence the expected coverage of $C_p(\boldsymbol{x})$ is $\int_{\{\boldsymbol{x} \in C_p\}} p(\boldsymbol{x} \,|\, \theta)\, d\boldsymbol{x} = p$, for all θ values. More generally, the frequentist coverage of all reference posterior p-credible regions in this example is exactly p.

Notice that, since the intrinsic loss $\delta\{\theta_0, \theta\}$ is invariant under one-to-one reparametrizations, the intrinsic statistic, the posterior intrinsic loss $d(\theta_0 \,|\, \boldsymbol{x})$, is also invariant so that, if $\phi(\theta)$ is a one-to-one transformation of θ, the intrinsic estimate of ϕ is $\phi^* = \phi(\theta^*)$, the intrinsic p-credible region of ϕ is $\phi(C_p)$, and the k-rejection region for ϕ is $\phi(R_k)$.

REFERENCES

Bernardo, J. M. (2004). Reference analysis. *Handbook of Statistics* **25** (D. Dey, ed.) Amsterdam: North-Holland, (to appear).

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Chichester: Wiley.

RÉSUMÉ

Une Vision Intégrée de la Statistique Mathématique: Construction Bayésienne Objective et Evaluation Fréquentielle

On enseigne souvent la statistique bayésienne, le cas échéant, séparément des méthodes conventionnelles. Il devient cependant clair que l'apparition de méthodes objectives bayésiennes puissantes (dont les résultats, ne dépend que du modèle adopté et des données observées) fournit une nouvelle perspective d'unification sur la plupart des méthodes établies et elles peuvent tre employées dans des situations où les méthodes fréquentielles ne peuvent l'être. D'autre part, les procédures fréquentielles fournissent des mécanismes d'évaluation et calibrent n'importe quelle méthode. Par conséquent, c'est peut-être l'heure à présent de considérer une approche intégrée sur l'statistique mathématique, où les méthodes bayésiennes objectives fournissent les éléments de construction inferentielle, et les méthodes fréquentielles les évaluations nécessaires.