

A BAYESIAN MATHEMATICAL STATISTICS PRIMER

José M. Bernardo

Universitat de València, Spain
jose.m.bernardo@uv.es

Bayesian Statistics is typically taught, if at all, after a prior exposure to frequentist statistics. It is argued that it may be appropriate to reverse this procedure. Indeed, 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 procedure. Hence, it may be the right time to consider an integrated approach to mathematical statistics, where objective Bayesian methods are first used to provide the building elements, and frequentist methods are then used to provide the necessary evaluation.

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 (indeed a recent survey of Bayesian papers indexed in the *Scientific Citation Index* shows an exponential growth), 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 usually invoke 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 pioneering landmark books by Jeffreys (1961), Lindley (1965), Zellner (1971), Press (1972) and Box and Tiao (1973), and *reference analysis*, whose development started in late 70's (see e.g. Bernardo Smith, 1994, §5.4, 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 methodology. And, indeed, it is not easy to combine 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.

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 $N(x|\mu, \sigma)$, with mean \bar{x} and standard deviation s , the interval $\bar{x} \pm t_{\alpha/2} s/\sqrt{n-1}$ 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 be 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 thus constructed would contain the true value of μ for $100(1 - \alpha)\%$ of the possible samples, thus providing a valuable calibration of the objective 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.

A particular implementation of an integrated programme in theoretical statistics along these lines is described below. This has been tested for three consecutive years in teaching a course on *Mathematical Statistics* (which is compulsory to all third year undergraduate students of both the degrees in *Mathematics* and in *Statistical Sciences*) at the *Universitat de València*, in 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 probability distributions are also discussed in this module, and they are used to introduce the important concept of the amount of information which the results from an experiment may be expected to provide. In particular, the *intrinsic discrepancy* between two probability distributions p_1 and p_2 for a random vector \mathbf{x} , defined as

$$\delta\{p_1, p_2\} = \min[k\{p_1 | p_2\}, k\{p_2 | p_1\}]$$

where $k\{p_j | p_i\}$ is the Kullback-Leibler directed divergence of p_j from p_i , defined by

$$k\{p_j | p_i\} = \int_{\mathbf{x}_i} p_i(\mathbf{x}) \log \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})} d\mathbf{x},$$

is shown to play an important rôle. The discrepancy between two probability families is defined as the minimum discrepancy between their elements. It immediately follows

that the intrinsic discrepancy between alternative models for the observed data \mathbf{x} is the minimum likelihood ratio for the true model, thus providing a useful natural calibration for this divergence measure.

The second module of the course is devoted to probability models. The concept of exchangeability, and the intuitive content of the 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 $\delta\{\Theta_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, which is the loss to be suffered from using a model in the family $\mathcal{M}_0 = \{p(\mathbf{x} | \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\theta}} \in \Theta_0, \tilde{\boldsymbol{\lambda}} \in \Lambda\}$ as a proxy for the assumed model $p(\mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\lambda})$, is defined as the intrinsic discrepancy $\delta\{p_{\mathbf{x}|\boldsymbol{\theta},\boldsymbol{\lambda}}, \mathcal{M}_0\}$ between the distribution $p(\mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\lambda})$ and the family of distributions in \mathcal{M}_0 , so that

$$\delta\{\Theta_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \inf_{\tilde{\boldsymbol{\theta}} \in \Theta_0, \tilde{\boldsymbol{\lambda}} \in \Lambda} \delta\{p_{\mathbf{x}|\boldsymbol{\theta},\boldsymbol{\lambda}}, p_{\mathbf{x}|\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\lambda}}}\}.$$

This intrinsic 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, a rather obvious requirement, which unfortunately many statistical methods fail to satisfy. For details, see Bernardo and Rueda (2002), and Bernardo (2005a, 2005b).

The last module of the course presents the frequentist paradigm as a set of methods designed to analyse the behaviour under repeated sampling of any proposed solution to a problem of statistical inference. In particular, these methods 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 evaluations are made using analytical techniques, when the relevant sampling distributions are easily derived, and Monte Carlo simulation techniques when they are not.

Theoretical expositions are completed with hands-on tutorials, where students are encouraged to analyse both real and simulated data at the computer lab using appropriate software.

AN EXAMPLE: EXPONENTIAL DATA

To illustrate the ideas proposed, we conclude by summarizing the details of a simple example, whose level of difficulty is typical of the course proposed.

Let \bar{x} be the mean of a random sample $\mathbf{x} = \{x_1, \dots, x_n\}$ from an exponential distribution

$$p(x | \theta) = \text{Ex}(x | \theta) = \theta e^{-x\theta}, \quad x > 0, \quad \theta > 0.$$

The (objective) reference prior in this problem is Jeffreys prior,

$$\pi(\theta) = \sqrt{i(\theta)} = \theta^{-1},$$

where

$$i(\theta) = - \int_{\mathcal{X}} p(x | \theta) \frac{\partial^2}{\partial \theta^2} \log p(x | \theta) dx$$

is Fisher information function. Using Bayes theorem, the corresponding reference posterior is found to be the Gamma distribution $\pi(\theta | \mathbf{x}) = \text{Ga}(\theta | n, n\bar{x})$, which only depends on the data \mathbf{x} through the sufficient statistic $\{\bar{x}, n\}$. For illustration, a random sample of size $n = 10$ was simulated from an exponential distribution with $\theta = 2$, which yielded $\bar{x} = 0.608$; the resulting reference posterior is represented in the lower panel of Figure 1.

The intrinsic loss is additive for independent observations. As a consequence, the loss from using model $p(\mathbf{x} | \tilde{\theta})$ as a proxy for $p(\mathbf{x} | \theta)$ (whose value is independent of the parametrization chosen) is $\delta\{\tilde{\theta}, \theta | n\} = n \delta_1\{\tilde{\theta}, \theta\}$, with

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

The reference posterior expectation of $\delta\{\tilde{\theta}, \theta | n\}$, which measures the expected discrepancy of $p(\mathbf{x} | \tilde{\theta})$ from the true model $p(\mathbf{x} | \theta)$, is the *intrinsic statistic function*

$$d(\tilde{\theta} | \mathbf{x}) = \int_0^\infty \delta\{\tilde{\theta}, \theta | n\} \pi(\theta | \mathbf{x}) d\theta,$$

whose exact value is represented in the top panel of Figure 1. The value $\theta^*(\mathbf{x})$ which minimizes $d(\tilde{\theta} | \mathbf{x})$ is the *Bayes estimator* which corresponds to the intrinsic discrepancy loss, or *intrinsic point estimator*. General results on (i) the invariance of $\delta\{\tilde{\theta}, \theta\}$ with respect to monotone transformations of the parameter, and (ii) the asymptotic normality of posterior distributions, yield the approximation

$$d(\tilde{\theta} | \mathbf{x}) \approx \frac{1}{2} \left[1 + n \delta\{\tilde{\theta}, \theta^*(\mathbf{x})\} \right], \quad \theta^*(\mathbf{x}) \approx \bar{x}^{-1} e^{-1/(2n)}.$$

Thus, the intrinsic estimator $\theta^*(\mathbf{x})$ is smaller than the mle $\hat{\theta}(\mathbf{x}) = \bar{x}^{-1}$. With the simulated data mentioned above, this is $\theta^* = 1.569$ represented in both panels of Figure 1 with a big dot. The approximation yields 1.565, and the mle is 1.645.

An *intrinsic p-credible region* is a p -credible region which contains points of *lowest posterior expected loss*. Hence, this is of the form

$$C_p \equiv \{\tilde{\theta}; d(\tilde{\theta} | \mathbf{x}) \leq k(p)\} \quad \text{and such that} \quad \int_{C(p)} \pi(\theta | \mathbf{x}) d\theta = p.$$

For instance, $C_{0.95}$ here consists of those parameter values with expected loss below 1.496, what yields the interval $C_{0.95} = [0.923, 2.658]$, shaded in the right panel of Figure 1. Moreover, the sampling distribution of \bar{x} (given θ and n) is $p(\bar{x} | \theta, n) = \text{Ga}(\bar{x} | n, n\theta)$, a Gamma distribution with mean θ^{-1} , and the *sampling* distribution of $t(\mathbf{x}) = \bar{x}\theta$ is $p(t | \theta, n) = \text{Ga}(t | n, n)$; but this is *also* the *posterior* distribution of $\phi(\theta) = \bar{x}\theta$, $\pi(\phi | \bar{x}, n) = \text{Ga}(\phi | n, n)$. Hence the expected frequentist coverage of the Bayesian p -credible region $C_p(\mathbf{x})$ is

$$\int_{\{\mathbf{x} \in C_p\}} p(\mathbf{x} | \theta) d\mathbf{x} = p, \quad \forall \theta > 0.$$

More generally, the frequentist coverage of *all* reference posterior p -credible regions in the exponential model is exactly p and, therefore, they are also exact frequentist confidence intervals for θ .

In a hypothesis testing situation, the *intrinsic k-rejection region* R_k consists of those $\tilde{\theta}$ values such that $d(\tilde{\theta} | \mathbf{x}) > k$, on the grounds that, given \mathbf{x} , the posterior expectation of the average log-likelihood ratio against them would be larger than k . For instance, with the data described (see the top panel of Figure 1), the values of $\tilde{\theta}$ smaller than 0.513 or larger than 4.771 yield values of the intrinsic statistic function larger than $k = \log(100) \approx 4.6$, and would therefore be rejected using this conventional threshold, since the average log-likelihood ratio against them is expected to be larger than $\log(100)$.

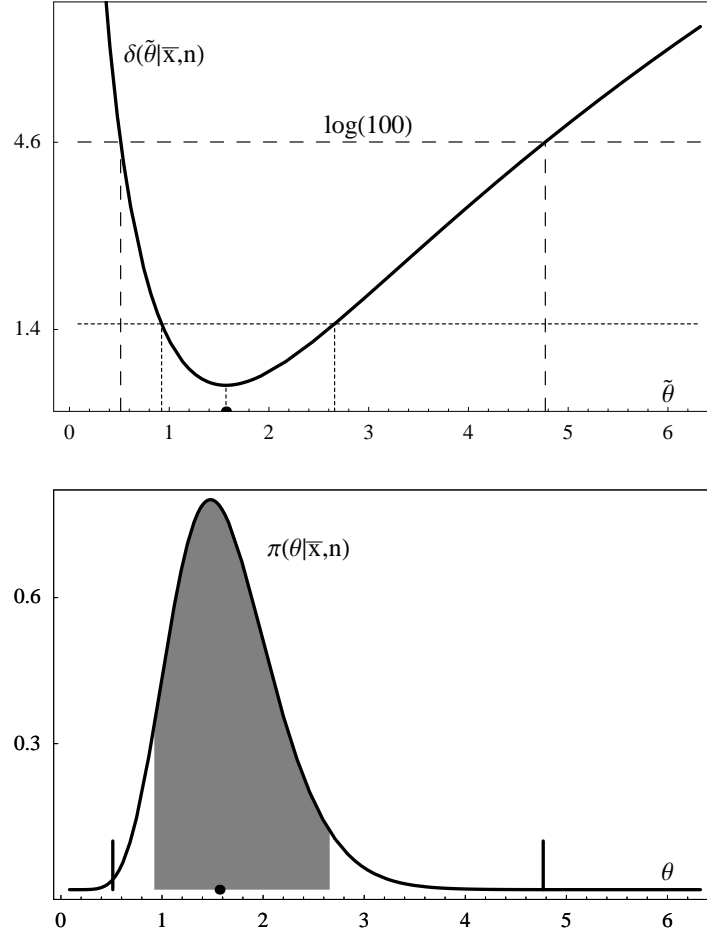


Figure 1. Intrinsic objective Bayesian inference for an exponential parameter

Notice that, since the intrinsic loss $\delta\{\tilde{\theta}, \theta\}$ is invariant under reparametrization, the intrinsic statistic function (the posterior intrinsic loss $d(\tilde{\theta} | \mathbf{x})$ from using $\tilde{\theta}$ instead of the true value of the parameter) is also invariant. Thus, 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 $\phi(R_k)$.

If prediction is desired, the reference (posterior) predictive distribution of a future observation x is

$$p(x | \mathbf{x}) = p(x | \bar{x}, n) = \int_0^{\infty} \text{Ex}(x | \theta) \text{Ga}(\theta | n, n\bar{x}) d\theta = \bar{x}^n \left(\frac{n}{\bar{x}n + x} \right)^{n+1},$$

which, as one would expect, converges to the true model $\text{Ex}(x | \theta)$ as $n \rightarrow \infty$. In particular, the (reference posterior) probability that a future observation x is larger than, say, t is

$$\Pr[x > t | \bar{x}, n] = \int_t^{\infty} p(x | \bar{x}, n) = \left(\frac{\bar{x}n}{\bar{x}n + t} \right)^n \quad (1)$$

which, as one would expect, converges to the true (conditional) predictive probability,

$$\Pr[x > t | \theta] = \int_t^{\infty} \text{Ex}(x | \theta) dx = e^{-t\theta}, \quad (2)$$

as $n \rightarrow \infty$. Notice, however, that the conventional *plug-in* predictive probability

$$\Pr[x > t | \bar{x}, n] \approx e^{-t\hat{\theta}},$$

which could be obtained by using in (2) some point estimate $\hat{\theta}$ of θ , may generally be very different from the correct value (1) and, hence, this would be seriously inappropriate for small sample sizes.

REFERENCES

- Bernardo, J. M. (2005a). Reference analysis. In D. K. Dey and C. R. Rao (Eds.), *Handbook of Statistics 25*, (pp.17–90). Amsterdam: Elsevier.
- Bernardo, J. M. (2005b). Intrinsic credible regions: An objective Bayesian approach to interval estimation (with discussion). *Test*, 14, 317–384.
- Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review*, 70, 351–372.
- Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley. (Second edition in preparation, 2006).
- Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.
- Jeffreys, H. (1961). *Theory of Probability* (3rd edition). Oxford: University Press.
- Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge: University Press.
- Press, S. J. (1972). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference* (2nd edition in 1982). Melbourne, FL: Krieger.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley. Reprinted in 1987, Melbourne, FL: Krieger.