Sociedad de Estadística
e Investigación Operativa

Test

Volume 14, Number 2. December 2005

Intrinsic Credible Regions: An Objective Bayesian Approach to Interval Estimation

José M. Bernardo
Departamento de Estadística e I. O.
Universidad de Valencia, Spain

Sociedad de Estadística e Investigación Operativa
Test (2005) *Vol.* **14**, *No.* 2, pp. 317–384

Sociedad de Estadística e Investigación Operativa
Test (2005) *Vol.* **14**, *No.* 2, pp. 317–384

Intrinsic Credible Regions: An Objective Bayesian Approach to Interval Estimation

José M. Bernardo[*]
Departamento de Estadística e I. O.
Universidad de Valencia, Spain

Abstract

This paper defines *intrinsic credible regions*, a method to produce objective Bayesian credible regions which only depends on the assumed model and the available data. *Lowest posterior loss* (LPL) regions are defined as Bayesian credible regions which contain values of minimum posterior expected loss; they depend both on the loss function and on the prior specification. An invariant, information-theory based loss function, the *intrinsic discrepancy*, is argued to be appropriate for scientific communication. Intrinsic credible regions are the lowest posterior loss regions with respect to the intrinsic discrepancy loss and the appropriate reference prior. The proposed procedure is completely general, and it is invariant under both reparametrization and marginalization. The exact derivation of intrinsic credible regions often requires numerical integration, but good analytical approximations are provided. Special attention is given to one-dimensional intrinsic credible intervals; their coverage properties show that they are always approximate (and sometimes exact) frequentist confidence intervals. The method is illustrated with a number of examples.

Key Words: Amount of information, intrinsic discrepancy, Bayesian asymptotics, confidence intervals, Fisher information, HPD regions, interval estimation, Jeffreys priors, LPL regions, objective priors, reference priors, point estimation, probability centred intervals, region estimation.

AMS subject classification: Primary 62F15; Secondary 62F25, 62B10.

1 Introduction and notation

This paper is mainly concerned with statistical inference problems such as occur in scientific investigation. Those problems are typically solved

[*] Correspondence to: José M. Bernardo. Departamento de Estadística e I.O., Facultad de Matemáticas, 46100-Burjassot, Spain. E-mail: jose.m.bernardo@uv.es

Work partially supported by grant MTM2004-05956 of MEC, Madrid, Spain

conditional on the assumption that a particular statistical model is an appropriate description of the probabilistic mechanism which has generated the data, and the choice of that model naturally involves an element of subjectivity. It has become standard practice however, to describe as "objective" any statistical analysis which only depends on the model assumed and the data observed. In this precise sense (and only in this sense) this paper provides an "objective" procedure to Bayesian region estimation.

Foundational arguments (Bernardo and Smith, 1994; de Finetti, 1970; Savage, 1954) dictate that scientists should elicit a unique (joint) prior distribution on all unknown elements of the problem on the basis of all available information, and use Bayes theorem to combine this with the information provided by the data, encapsulated in the likelihood function, to obtain a joint posterior distribution. Standard probability theory may then be used to derive from this joint posterior the posterior distribution of the quantity of interest; mathematically this is the final result of the statistical analysis. Unfortunately however, elicitation of the joint prior is a formidable task, specially in realistic models with many nuisance parameters which rarely have a simple interpretation, or in scientific inference, where some sort of *consensus* on the elicited prior would obviously be required. In this context, the (unfortunately very frequent) naïve use of simple proper "flat" priors (often a limiting form of a conjugate family) as presumed "noninformative" priors often hides important unwarranted assumptions which may easily dominate, or even invalidate, the analysis: see *e.g.*, Berger (2000), and references therein. The uncritical (ab)use of such "flat" priors should be strongly discouraged. An appropriate *reference prior* (Berger and Bernardo, 1992c; Bernardo, 1979b, 2005b) should instead be used.

As mentioned above, from a Bayesian viewpoint, the final outcome of a problem of inference about *any* unknown quantity is simply the posterior distribution of that quantity. Thus, given some data $\boldsymbol{x}$ and conditions C, *all* that can be said about any function $\boldsymbol{\theta}(\boldsymbol{\omega})$ of the parameter vector $\boldsymbol{\omega}$ which govern the model is contained in the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{x}, C)$, and *all* that can be said about some function $\boldsymbol{y}$ of future observations from the same model is contained in its posterior predictive distribution $p(\boldsymbol{y} \mid \boldsymbol{x}, C)$. Indeed (Bernardo, 1979a), Bayesian inference is a decision problem where the action space is the class of those posterior probability distributions of the quantity of interest which are compatible with accepted assumptions.

However, to make it easier for the user to assimilate the appropriate conclusions, it is often convenient to *summarize* the information contained in the posterior distribution, while retaining as much of the information as possible. This is conveniently done by providing sets of possible values of the quantity of interest which, in the light of the data, are likely to be "close" to its true value. The pragmatic importance of these *region estimates* should not be underestimated; see Guttman (1970), Blyth (1986), Efron (1987), Hahn and Meeker (1991), Burdick and Graybill (1992), Eberly and Casella (2003), and references therein, for some monographic works on this topic. In this paper, a new objective Bayesian solution to this *region estimation* problem is proposed and analyzed.

1.1 Notation

It will be assumed that probability distributions may be described through their probability density functions, and no notational distinction will be made between a random quantity and the particular values that it may take. Bold italic roman fonts are used for observable random vectors (typically data) and bold italic greek fonts for unobservable random vectors (typically parameters); lower case is used for variables and upper case calligraphic for their dominion sets. Moreover, the standard mathematical convention of referring to functions, say $f_{\boldsymbol{x}}$ and $g_{\boldsymbol{x}}$ of $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, respectively by $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ will be used throughout. Thus, the conditional probability density of data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ given $\boldsymbol{\theta}$ will be represented by either $p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}$ or $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$, with $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) \geq 0$ and $\int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} = 1$, and the posterior distribution of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ given $\boldsymbol{x}$ will be represented by either $p_{\boldsymbol{\theta}\,|\,\boldsymbol{x}}$ or $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})$, with $p(\boldsymbol{\theta}\,|\,\boldsymbol{x}) \geq 0$ and $\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\,\mathrm{d}\boldsymbol{\theta} = 1$. This admittedly imprecise notation will greatly simplify the exposition. If the random vectors are discrete, these functions naturally become probability mass functions, and integrals over their values become sums. Density functions of specific distributions are denoted by appropriate names. Thus, if x is an observable random variable with a normal distribution of mean μ and variance σ^2, its probability density function will be denoted $\mathrm{N}(x\,|\,\mu, \sigma)$. If the posterior distribution of μ is Student with location $\overline{x}$, scale s, and n degrees of freedom, its probability density function will be denoted $\mathrm{St}(\mu\,|\,\overline{x}, s, n)$.

1.2 Problem statement

The argument is always defined in terms of some *parametric model* of the general form $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, which describes the conditions under which data have been generated. Thus, data $\boldsymbol{x}$ are assumed to consist of one observation of the random vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, with probability density $p(\boldsymbol{x} \,|\, \boldsymbol{\omega})$, for some $\boldsymbol{\omega} \in \boldsymbol{\Omega}$. Often, but not necessarily, data will consist of a random sample $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of fixed size n from some distribution with, say, density $p(\boldsymbol{y} \,|\, \boldsymbol{\omega})$, $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$, in which case $p(\boldsymbol{x} \,|\, \boldsymbol{\omega}) = \prod_{j=1}^{n} p(\boldsymbol{y}_j \,|\, \boldsymbol{\omega})$, and $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Y}}^n$.

Let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ be some vector of interest; without loss of generality, the assumed model $\mathcal{M}$ may be reparametrized in the form

$$\mathcal{M} \equiv \{\, p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda} \,\}, \tag{1.1}$$

where $\boldsymbol{\lambda}$ is some vector of nuisance parameters; this is often simply referred to as "model" $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$. Conditional on the assumed model, all valid Bayesian inferential statements about the value of $\boldsymbol{\theta}$ are encapsulated in its posterior distribution

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto \int_{\boldsymbol{\Lambda}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\, p(\boldsymbol{\theta}, \boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\lambda}, \tag{1.2}$$

which *combines* the information provided by the data $\boldsymbol{x}$ with any other information about $\boldsymbol{\theta}$ contained in the prior density $p(\boldsymbol{\theta}, \boldsymbol{\lambda})$.

With no commonly agreed prior information on $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ the *reference prior function* for the quantity of interest, a mathematical description of that situation which maximizes the missing information about the quantity of interest $\boldsymbol{\theta}$ which will be denoted by $\pi(\boldsymbol{\theta})\, \pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})$, should be used to obtain the corresponding *reference posterior*,

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto \pi(\boldsymbol{\theta}) \int_{\boldsymbol{\Lambda}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\lambda}. \tag{1.3}$$

To describe the inferential content of the posterior distribution $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ of the quantity of interest and, in particular, that of the reference posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, it is often convenient to quote regions $R \subset \boldsymbol{\Theta}$ of given (posterior) probability under $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, often called credible regions.

This paper concentrates on credible regions for parameter values. However, the ideas may be extended to prediction problems by using the *pos-*

terior predictive density of the quantity $\boldsymbol{y}$ to be predicted, namely $p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \int_\Omega p(\boldsymbol{y}\,|\,\boldsymbol{\omega})\; p(\boldsymbol{\omega}\,|\,\boldsymbol{x})\,\mathrm{d}\boldsymbol{\omega}$, in place of the posterior density of $\boldsymbol{\theta}$.

Definition 1.1 (Credible region). *A* (posterior) q-credible region *for* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *is a subset* $R_q(\boldsymbol{x}, \boldsymbol{\Theta})$ *of the parameter space* $\boldsymbol{\Theta}$ *such that,*

$$R_q(\boldsymbol{x}, \boldsymbol{\Theta}) \subset \boldsymbol{\Theta}, \quad \int_{R_q(\boldsymbol{x}, \boldsymbol{\Theta})} p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\, d\boldsymbol{\theta} = q, \quad 0 < q \leq 1.$$

Thus, given data $\boldsymbol{x}$*, the true value of* $\boldsymbol{\theta}$ *belongs to* $R_q(\boldsymbol{x}, \boldsymbol{\Theta})$ *with (posterior) probability* q*.*

If there is no danger of confusion, dependence on available data $\boldsymbol{x}$ and explicit mention of the parametrization used will both be dropped from the notation, and a q-credible region $R_q(\boldsymbol{x}, \boldsymbol{\Theta})$ will simply be denoted by R_q.

Credible regions are invariant under reparametrization. Thus, for any q-credible region $R_q(\boldsymbol{x}, \boldsymbol{\Theta})$ for $\boldsymbol{\theta}$ and for any one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta}) \in \phi(\boldsymbol{\Theta}) = \boldsymbol{\Phi}$ of the parameter $\boldsymbol{\theta}$, $R_q(\boldsymbol{x}, \boldsymbol{\Phi}) = \boldsymbol{\phi}\{R_q(\boldsymbol{x}, \boldsymbol{\Theta})\}$ is a q-credible region for $\boldsymbol{\phi}$. However, for any given q there are generally *infinitely many* credible regions. Many efforts have been devoted to the selection of an appropriate credible region.

Sometimes, credible regions are selected to have minimum size (length, area, volume), resulting in *highest posterior density* (HPD) regions, where all points in the region have larger posterior probability density than all points outside. However, HPD regions are *not* invariant under reparametrization: the image $R_q(\boldsymbol{x}, \boldsymbol{\Phi}) = \boldsymbol{\phi}\{R_q(\boldsymbol{x}, \boldsymbol{\Theta})\}$ of a HPD q-credible region for $\boldsymbol{\theta}$ will be a q-credible region for $\boldsymbol{\phi}$, but will not generally be HPD. Thus, the apparently intuitive idea behind the definition of HPD regions is found to be illusory, for it totally depends on the (arbitrary) parametrization chosen to describe the problem.

In *one dimensional problems*, posterior quantiles are often used as an alternative to HPD regions to specify credible regions. Thus, if $\theta_q = \theta_q(\boldsymbol{x})$ is the posterior q-quantile of θ, then $R_q(\boldsymbol{x}, \Theta) = \{\theta;\, \theta \leq \theta_q\}$ is a one-sided, typically unique q-credible interval, and it is invariant under reparametrization. Posterior quantiles may be used to define *probability centred* q-credible intervals of the form

$$R_q(\boldsymbol{x}, \Theta) = \{\theta;\; \theta_{(1-q)/2} \leq \theta \leq \theta_{(1+q)/2}\},$$

so that there is the same probability, namely $(1-q)/2$, that the true value of θ is at either side of the interval. Probability centred intervals are easier to compute, and they are often quoted in preference to HPD regions. However, probability centred credible intervals are only really appealing when the posterior density has a unique *interior* mode and, moreover, they have a crucial limitation: they are not uniquely defined in problems with more than one dimension.

Example 1.1 (Credible intervals for a binomial parameter). *Consider a set $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of n independent Bernoulli observations with parameter $\theta \in \Theta = (0, 1)$, so that $p(x \,|\, \theta) = \theta^x(1-\theta)^{1-x}$, and the likelihood function is $p(\boldsymbol{x} \,|\, \theta) = \theta^r(1-\theta)^{n-r}$, with $r = \sum_{j=1}^n x_j$. The reference prior, which in this case is also Jeffreys prior, is $\pi(\theta) = \mathrm{Be}(\theta \,|\, \frac{1}{2}, \frac{1}{2})$, and the reference posterior is $\pi(\theta \,|\, r, n) = \mathrm{Be}(\theta \,|\, r+\frac{1}{2}, n-r+\frac{1}{2}) \propto \theta^{r-1/2}(1-\theta)^{n-r-1/2}$. A (posterior) q-credible region for θ is any subset of R_q of $(0,1)$ such that $\int_{R_q} \mathrm{Be}(\theta \,|\, r+\frac{1}{2}, n-r+\frac{1}{2})\, d\theta = q$.*

Consider now the one-to-one (variance stabilizing) reparametrization $\phi = 2 \arcsin \sqrt{\theta}$, $\phi \in \Phi = (0, \pi)$, so that $\theta = \sin^2(\phi/2)$. Changing variables, the reference posterior density of ϕ is

$$\pi(\phi \,|\, r, n) = \frac{\pi(\theta \,|\, r, n)}{|\partial\phi(\theta)/\partial\theta|}\bigg|_{\theta=\sin^2(\phi/2)} \propto (\sin^2[\phi/2])^r(\cos^2[\phi/2])^{n-r}, \tag{1.4}$$

which conveys precisely the same information that $\pi(\theta \,|\, r, n)$. Clearly, if the set $R_q(r, n, \Theta)$ is a q-credible region for θ then $R_q(r, n, \Phi) = \phi\{R_q(r, n, \Theta)\}$ will be a q-credible region for ϕ; however, if $R_q(r, n, \Theta)$ is HPD *for θ, then $R_q(r, n, \Phi)$ will generally not be* HPD *for ϕ.*

For a numerical illustration, consider the case $n = 10$, $r = 2$, so that the reference posterior is the beta density $\mathrm{Be}(\theta \,|\, 2.5, 8.5)$ *represented in the left panel of Figure 1. Numerical integration or the use of the incomplete beta integral shows that the* 0.95 HPD *credible interval is the set* (0.023, 0.462) *of those θ values whose posterior density is larger than* 0.585 *(shaded region in that figure). The reference posterior of ϕ, given by Equation (1.4), is shown on the right panel of Figure 1; the θ-*HPD *interval transforms into* $\phi[(0.023,\ 0.462)] = (0.308,\ 1.495)$ *which is a* 0.95*-credible interval for ϕ, but clearly* not HPD. *The* 0.95 *probability centred credible interval for θ is* (0.044, 0.503)*, slightly to the right of the* HPD *interval. Consider now the case $n = 10$, $r = 0$, so that no successes have been observed in ten trials. The reference posteriors densities of θ and ϕ are now both monotone*

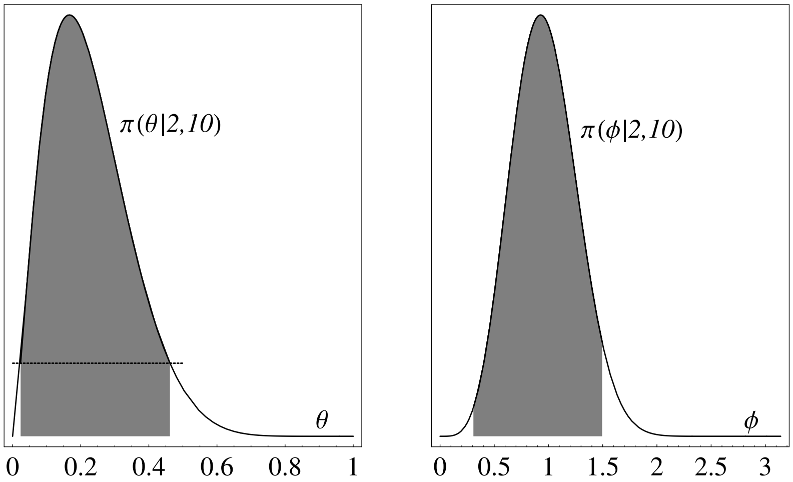

Figure 1: HPD *credible regions do not remain* HPD *under reparametrization.*

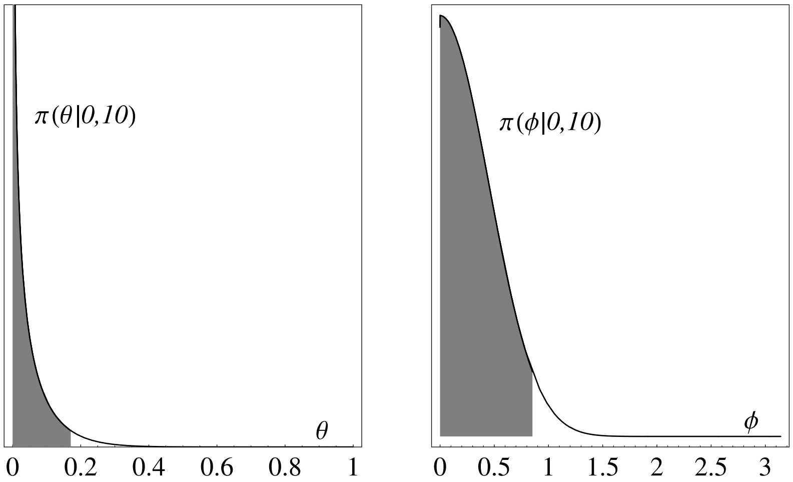

Figure 2: *Probability centred credible intervals are not appropriate if posteriors have not a unique interior mode.*

decreasing from zero (see Figure 2). The HPD *interval for* θ *is* $(0, 0.170)$*; this transforms into* $\phi[(0,\, 0.170)] = (0,\, 0.852)$, *which now is also* HPD *in* ϕ*. Clearly, probability centred intervals do not make much intuitive sense in this case, for they would leave out the neighbourhood of zero, which is by far the region more likely to contain the true parameter value.*

Conventional frequentist theory fails to produce a convincing confidence interval in this (very simple) example. Indeed, since data are discrete, an exact non-randomized confidence interval of level $1-q$ *does not exist for most q-values. On the other hand the frequentist coverage of (exact) objective q-credible intervals may generally be shown to be* $q+O(n^{-1})$*; thus, Bayesian q-credible regions typically produce* approximate *confidence intervals of level* $1-q$*. See Section 5 for further discussion.*

As the preceding example illustrates, even in simple one-dimensional problems, there is no generally agreed solution on the appropriate choice of credible regions. As one would expect, the situation only gets worse in many dimensions.

In the next section, a decision theory argument is used to propose a new procedure for the selection of credible intervals, a procedure designed to overcome the problems discussed above.

2 Lowest posterior loss (LPL) credible regions

Let $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ be some vector of interest and suppose that available data $\boldsymbol{x}$ are assumed to consist of one observation from

$$\mathcal{M}\equiv\{\,p(\boldsymbol{x}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda}),\ \boldsymbol{x}\in\boldsymbol{\mathcal{X}},\ \boldsymbol{\theta}\in\boldsymbol{\Theta},\ \boldsymbol{\lambda}\in\boldsymbol{\Lambda}\,\},$$

where $\boldsymbol{\lambda}\in\boldsymbol{\Lambda}$ is some vector of nuisance parameters. Let $p(\boldsymbol{\theta},\boldsymbol{\lambda})$ the the joint prior for $(\boldsymbol{\theta},\boldsymbol{\lambda})$, let $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\propto\int_{\boldsymbol{\Lambda}}\,p(\boldsymbol{x}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda})\;p(\boldsymbol{\theta},\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda}$ be the corresponding marginal posterior for $\boldsymbol{\theta}$, and let $\ell\{\boldsymbol{\theta}_0,\boldsymbol{\theta}\}$ the *loss* to be suffered if a particular value $\boldsymbol{\theta}_0\in\boldsymbol{\Theta}$ of the parameter were used as a proxy for the unknown true value of $\boldsymbol{\theta}$ in the specific application under consideration. The expected loss from using $\boldsymbol{\theta}_0$ is then

$$l\{\boldsymbol{\theta}_0\,|\,\boldsymbol{x}\}=\mathrm{E}_{\boldsymbol{\theta}\,|\,\boldsymbol{x}}[\ell\{\boldsymbol{\theta}_0,\boldsymbol{\theta}\}]=\int_{\boldsymbol{\Theta}}\ell\{\boldsymbol{\theta}_0,\boldsymbol{\theta}\}\,p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\,\mathrm{d}\boldsymbol{\theta}\qquad(2.1)$$

and the optimal (Bayes) estimate of $\boldsymbol{\theta}$ with respect to this loss (given the assumed prior), is

$$\boldsymbol{\theta}^*(\boldsymbol{x})=\arg\inf_{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}}l\{\boldsymbol{\theta}_0\,|\,\boldsymbol{x}\}.\qquad(2.2)$$

As mentioned before, with no commonly agreed prior information on $(\boldsymbol{\theta},\boldsymbol{\omega})$ the prior $p(\boldsymbol{\theta},\boldsymbol{\lambda})$ will typically be taken to be the reference prior function for the quantity of interest, $\pi(\boldsymbol{\theta})\,\pi(\boldsymbol{\omega}\,|\,\boldsymbol{\theta})$.

More generally, the loss to be suffered if $\boldsymbol{\theta}_0$ were used as a proxy for $\boldsymbol{\theta}$ could also depend on the true value of the nuisance parameter $\boldsymbol{\lambda}$. In this case, the loss function would be of the general form $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ and the expected loss from using $\boldsymbol{\theta}_0$ would be

$$l\{\boldsymbol{\theta}_0 \,|\, \boldsymbol{x}\} = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, p(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\lambda}, \tag{2.3}$$

where $p(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\, p(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the joint posterior of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$.

With a loss structure precisely defined, coherence *dictates* that parameter values with smaller expected loss should always be preferred. For reasonable loss functions, a typically unique credible region may be selected as a *lowest posterior loss* (LPL) region, where all points in the region have smaller posterior expected loss than all points outside.

Definition 2.1 (Lowest posterior loss credible region). *Let data $\boldsymbol{x}$ consist of one observation from $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, and let $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ be the loss to be suffered if $\boldsymbol{\theta}_0$ were used as a proxy for $\boldsymbol{\theta}$. A* lowest posterior loss q-credible region *is a subset $R_q^\ell = R_q^\ell(\boldsymbol{x}, \boldsymbol{\Theta})$ of the parameter space $\boldsymbol{\Theta}$ such that,*

(i). $\int_{R_q^\ell} p(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} = q,$

(ii). $\forall \boldsymbol{\theta}_i \in R_q^\ell,\ \forall \boldsymbol{\theta}_j \notin R_q^\ell,\ l(\boldsymbol{\theta}_i \,|\, \boldsymbol{x}) \le l(\boldsymbol{\theta}_j \,|\, \boldsymbol{x}),$

where $l(\boldsymbol{\theta}_i \,|\, \boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \ell\{\boldsymbol{\theta}_i,\, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, p(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}$.

Lowest posterior loss regions obviously depend on the particular loss function used. In principle, any loss function could be used. However, in scientific inference one would expect the loss function to be *invariant* under one-to-one reparametrization. Indeed, if θ is a positive quantity of interest, the loss suffered from using θ_0 instead of the true value of θ should be precisely the same the same as, say, the loss suffered from using $\log \theta_0$ instead of $\log \theta$. Moreover, the (arbitrary) parameter is only a label for the model. Thus, for any one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ in $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\boldsymbol{\Theta})$, the model $\{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is precisely the same as the (reparametrized) model $\{p(\boldsymbol{x} \,|\, \boldsymbol{\phi}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$; the conclusions to be derived from available data $\boldsymbol{x}$ should be precisely the same whether one chooses to work in terms of $\boldsymbol{\theta}$ or in terms of $\boldsymbol{\phi}$. Thus, in scientific inference, where only truth is supposed to matter, the loss suffered $\ell\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\}$ from using $\boldsymbol{\theta}_0$ instead of $\boldsymbol{\theta}$ should *not* measure the (irrelevant) discrepancy in the parameter space $\boldsymbol{\Theta}$ between the *parameter values* $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$, but the (relevant) discrepancy in the appropriate functional space between the *models* $p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}_0}$ and $p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}$ which they label. Such a loss function, of general form $\ell\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\} = \ell\{p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}_0}, p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}\}$, will obviously be invariant under one-to-one reparametrizations, so that for any such transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$, one will have $\ell\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\} = \ell\{\boldsymbol{\phi}_0, \boldsymbol{\phi}\}$, with $\boldsymbol{\phi}_0 = \boldsymbol{\phi}(\boldsymbol{\theta}_0)$, as required.

Loss functions which depend on the models they label rather than on the parameters themselves are known as *intrinsic* loss functions (Robert, 1996). This concept is *not* related to the concepts of "intrinsic Bayes factors" and "intrinsic priors" introduced by Berger and Pericchi (1996).

Definition 2.2 (Intrinsic loss function). *Consider the probability model* $\mathcal{M} \equiv \{\, p(\boldsymbol{x}\,|\,\boldsymbol{\omega}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\,\}$. *An* intrinsic loss function *for* $\mathcal{M}$ *is a symmetric, non-negative function* $\ell\{\boldsymbol{\omega}_0, \boldsymbol{\omega}\}$ *of the general form*

$$\ell\{\boldsymbol{\omega}_0, \boldsymbol{\omega}\} = \ell\{\boldsymbol{\omega}, \boldsymbol{\omega}_0\} = \ell\{p_{\boldsymbol{x}\,|\,\boldsymbol{w}_0}, p_{\boldsymbol{x}\,|\,\boldsymbol{w}}\}$$

which is zero if, and only if, $p(\boldsymbol{x}\,|\,\boldsymbol{\omega}_0) = p(\boldsymbol{x}\,|\,\boldsymbol{\omega})$ *almost everywhere.*

Well known examples of intrinsic loss functions include the $\mathcal{L}_1$ norm,

$$\ell_1\{\boldsymbol{\omega}_0, \boldsymbol{\omega}\} = \int_{\mathcal{X}} |p(\boldsymbol{x}\,|\,\boldsymbol{\omega}_0) - p(\boldsymbol{x}\,|\,\boldsymbol{\omega})|\, \mathrm{d}\boldsymbol{x} \tag{2.4}$$

and the $\mathcal{L}_\infty$ norm

$$\ell_\infty\{\boldsymbol{\omega}_0, \boldsymbol{\omega}\} = \sup_{\boldsymbol{x} \in \mathcal{X}} |p(\boldsymbol{x}\,|\,\boldsymbol{\omega}_0) - p(\boldsymbol{x}\,|\,\boldsymbol{\omega})|. \tag{2.5}$$

All intrinsic loss functions are invariant under reparametrization, but they they are not necessarily invariant under one-to-one transformations of $\boldsymbol{x}$. Thus, ℓ_1 in Equation (2.4) is invariant in this sense, but ℓ_∞ in Equation (2.5) is *not*. Intrinsic loss functions which are invariant under one-to-one transformations of the data are typically also invariant under reduction to sufficient statistics. For example, if $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x}) \in \mathcal{T}$ is sufficient for the model under consideration, so that $p(\boldsymbol{x}\,|\,\boldsymbol{\omega}) = p(\boldsymbol{t}\,|\,\boldsymbol{\omega})\, p(\boldsymbol{s}\,|\,\boldsymbol{t})$, where $\boldsymbol{s} = \boldsymbol{s}(\boldsymbol{x})$ is an ancillary statistic, the intrinsic ℓ_1 loss becomes

$$\begin{aligned}\ell_1\{\boldsymbol{\omega}_0, \boldsymbol{\omega}\} &= \int_{\mathcal{X}} p(\boldsymbol{x} \,|\, \boldsymbol{\omega}) \left| \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}_0)}{p(\boldsymbol{x} \,|\, \boldsymbol{\omega})} - 1 \right| \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathcal{T}} p(\boldsymbol{t} \,|\, \boldsymbol{\omega}) \left| \frac{p(\boldsymbol{t} \,|\, \boldsymbol{\omega}_0)}{p(\boldsymbol{t} \,|\, \boldsymbol{\omega})} - 1 \right| \mathrm{d}\boldsymbol{t} \\ &= \int_{\mathcal{T}} |p(\boldsymbol{t} \,|\, \boldsymbol{\omega}_0) - p(\boldsymbol{t} \,|\, \boldsymbol{\omega})| \, \mathrm{d}\boldsymbol{t}.\end{aligned}$$

Hence, the ℓ_1 loss would be the same whether one uses the full model $p(\boldsymbol{x} \,|\, \boldsymbol{\omega})$ or the marginal model $p(\boldsymbol{t} \,|\, \boldsymbol{\omega})$ induced by the sampling distribution of the sufficient statistic $\boldsymbol{t}$. The loss ℓ_∞ however is *not* invariant is this statistically important sense.

The conclusions to be derived from any data set $\boldsymbol{x}$ should obviously the same as those derived from reduction to any sufficient statistic; hence, only intrinsic loss functions which are invariant under reduction by sufficiency should really be considered.

Example 2.1 (Credible intervals for a binomial parameter (continued)). *Consider again the problem considered in Example 1.1 and take the ℓ_1 loss function of Equation (2.4). Since this loss is invariant under reduction to a sufficient statistic, the expected loss from using θ_0 rather than θ may be found using the sampling distribution $p(r \,|\, \theta) = \mathrm{Bi}(r \,|\, n, \theta)$ of the sufficient statistic r. This yields*

$$\begin{aligned}l_1\{\theta_0 \,|\, r, n\} &= \int_0^1 \ell_1\{\theta_0, \theta\} \, \mathrm{Be}(\theta \,|\, r + \tfrac{1}{2}, n - r + \tfrac{1}{2}) \, d\theta \\ \ell_1\{\theta_0, \theta\} &= \sum_{r=0}^{n} |\mathrm{Bi}(r \,|\, n, \theta_0) - \mathrm{Bi}(r \,|\, n, \theta)|.\end{aligned}$$

The expected loss $l_1\{\theta_0 \,|\, r, n\}$ is shown in the upper panel of Figure 3 for the case $r = 2$ and $n = 10$ discussed before. This has a unique minimum at $\theta^* = 0.210$ which is therefore the Bayes estimator for this loss (marked with a solid dot in the lower panel of Figure 3). The 0.95-LPL *credible interval for this loss is numerically found to consist of the set* $(0.037, 0.482)$ *whose expected loss is lower than* 1.207 *(shaded region in the lower panel of Figure 3). Since intrinsic loss functions are invariant under reparametrizations, the Bayes estimate ϕ^* and* LPL *q-credible region of some one-to-one*

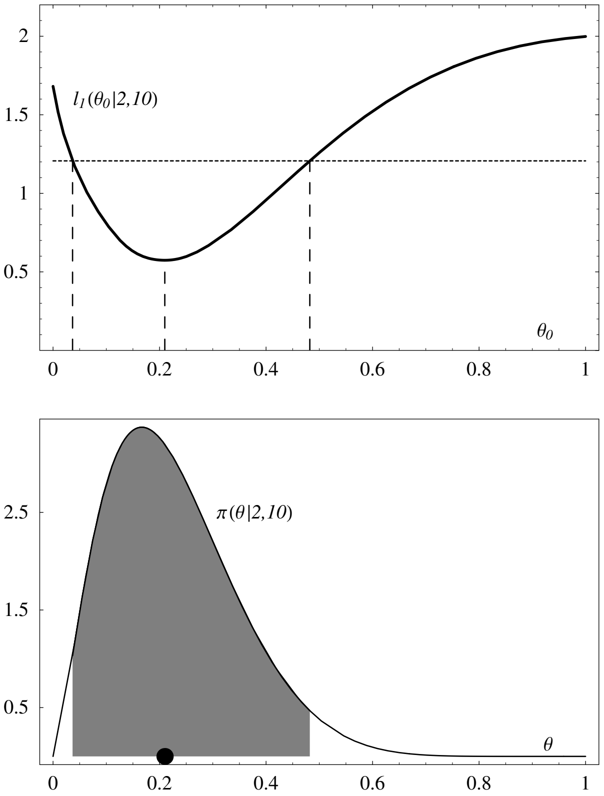

Figure 3: *Bayes estimator and* LPL 0.95*-credible region for a binomial parameter using the* $\mathcal{L}_1$ *intrinsic loss.*

function of ϕ *will simply be* $\phi(\theta^*)$ *and* $\phi[R_q^\ell(r, n, \Theta)]$. *For the variance-stabilizing transformation* $\phi(\theta) = 2 \arcsin \sqrt{\theta}$ *already considered in Example* 1.1 *these are, respectively,* 0.952 *and* (0.385, 1.535).

Notice that if one were to use a conventional, not invariant loss function, the results would not *be invariant under reparametrization. For instance, with a quadratic loss* $\ell\{\theta_0, \theta\} = (\theta_0 - \theta)^2$, *the Bayes estimator is the posterior mean,* $E[\theta \,|\, r, n] = 0.227$*; similarly, the Bayes estimator for* ϕ *would be its posterior mean* $E[\phi \,|\, r, n] = 0.965$, *which is different from* $\phi(0.227) = 0.994$*; credible regions would be similarly inconsistent. Yet, it would be hard to argue, say to a quality engineer, that your best guess for the proportion of defective items is* θ^*, *but that your best guess for* $\log \theta$ *is* not $\log \theta^*$.

In the next section, a particular, invariant intrinsic loss, the *intrinsic discrepancy* will be introduced. It is argued that this provides a far better conventional loss function of choice for mathematical statistics than the ubiquitous, overused quadratic loss.

3 The intrinsic discrepancy loss

Probability theory makes frequent use of *divergence measures* between probability distributions. The total variation distance, Hellinger distance, Kullback-Leibler logarithmic divergence, and Jeffreys logarithmic divergence are all frequently cited; see, for example, Kullback (1968), and Gutiérrez-Peña (1992) for precise definitions and properties. Each of those divergence measures may be used to define a type of convergence. It has been found, however, that the behaviour of many important limiting processes, in both probability theory and statistical inference, is better described in terms of another information-theory related divergence measure, the *intrinsic discrepancy* (Bernardo and Rueda, 2002), which is now defined and illustrated.

Definition 3.1 (Intrinsic discrepancy). *Consider two probability distributions of a random vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, specified by their density functions $p_1(\boldsymbol{x})$, $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_1 \subset \boldsymbol{\mathcal{X}}$, and $p_2(\boldsymbol{x})$, $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_2 \subset \boldsymbol{\mathcal{X}}$, with either identical or nested supports. The* intrinsic discrepancy $\delta\{p_1, p_2\}$ *between p_1 and p_2 is*

$$\delta\{p_1, p_2\} = \min\left\{ \int_{\boldsymbol{\mathcal{X}}_1} p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\, d\boldsymbol{x},\ \int_{\boldsymbol{\mathcal{X}}_2} p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{x})}{p_1(\boldsymbol{x})}\, d\boldsymbol{x} \right\}, \quad (3.1)$$

provided one of the integrals (or sums) is finite. The intrinsic discrepancy $\delta\{\mathcal{F}_1, \mathcal{F}_2\}$ between two families $\mathcal{F}_1$ and $\mathcal{F}_2$ of probability distributions is the minimum intrinsic discrepancy between their elements,

$$\delta\{\mathcal{F}_1, \mathcal{F}_2\} = \inf_{p_1 \in \mathcal{F}_1,\, p_2 \in \mathcal{F}_2} \delta\{p_1,\, p_2\}. \quad (3.2)$$

It is immediate from Definition 3.1 that the intrinsic discrepancy between two probability distributions may be written in terms of their two possible *directed divergences* (Kullback and Leibler, 1951) as

$$\delta\{p_2, p_1\} = \min\{\ \kappa\{p_2 \,|\, p_1\},\ \kappa\{p_1 \,|\, p_2\}\ \} \quad (3.3)$$

where the $\kappa\{p_j \,|\, p_i\}$'s are the non-negative quantities defined by

$$\kappa\{p_j \,|\, p_i\} = \int_{\boldsymbol{\mathcal{X}}_i} p_i(\boldsymbol{x}) \log \frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x}, \quad \text{with} \ \ \boldsymbol{\mathcal{X}}_i \subseteq \boldsymbol{\mathcal{X}}_j. \quad (3.4)$$

which are invariant under one-to-one transformations of $\boldsymbol{x}$. Since $\kappa\{p_j \,|\, p_i\}$ is the expected value of the logarithm of the density (or probability) ratio for p_i against p_j when p_i is true, it also follows from Definition 3.1 that, if p_1 and p_2 describe two alternative models, one of which is assumed to generate the data, their intrinsic discrepancy $\delta\{p_1, p_2\}$ is the *minimum expected log-likelihood ratio in favour of the model which generates the data* (the "true" model).

The intrinsic discrepancy $\delta\{p_1, p_2\}$ is a divergence measure (*i.e.*, it is symmetric, non-negative and zero iff $p_1 = p_2$ a.e.) with two added important properties which make it virtually unique: (i) the intrinsic discrepancy is still defined when the supports are strictly nested; hence, the intrinsic discrepancy $\delta\{p, \hat{p}\}$ between, say a distribution p with support on $\mathbb{R}$ and its approximation $\hat{p}$ with support on some compact subset $[a, b]$ may be computed; and (ii) the intrinsic discrepancy is additive for independent observations. As a consequence of (ii), the intrinsic discrepancy $\delta\{\theta_1, \theta_2\}$ between two possible joint models $\prod_{j=1}^n p_1(x_j \,|\, \theta_1)$ and $\prod_{j=1}^n p_2(x_j \,|\, \theta_2)$ for a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is simply n times the discrepancy between $p_1(x \,|\, \theta_1)$ and $p_2(x \,|\, \theta_2)$.

Theorem 3.1 (Properties of the intrinsic discrepancy). *Let p_1 and p_2 be any two probability densities for the random vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ with either identical or nested supports $\boldsymbol{\mathcal{X}}_1$ and $\boldsymbol{\mathcal{X}}_2$. Their intrinsic discrepancy $\delta\{p_1, p_2\}$ is*

(i). Symmetric: $\delta\{p_1, p_2\} = \delta\{p_2, p_1\}$

(ii). Non-negative: $\delta\{p_1, p_2\} \geq 0$, *and*
$\delta\{p_1, p_2\} = 0$ *if, and only if,* $p_1(\boldsymbol{x}) = p_2(\boldsymbol{x})$ *a.e.*

(iii). Defined for strictly nested supports:
if $\boldsymbol{\mathcal{X}}_i \subset \boldsymbol{\mathcal{X}}_j$, *then* $\delta\{p_i, p_j\} = \delta\{p_j, p_i\} = \kappa\{p_j \,|\, p_i\}$.

(iv). Invariant: If $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{x})$ is one-to-one and $q_i(\boldsymbol{z})$ is the probability density of $\boldsymbol{z}$ induced by $p_i(\boldsymbol{x})$, then $\delta\{p_1, p_2\} = \delta\{q_1, q_2\}$

(v). Additive for independent observations: If $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, and
$p_i(\boldsymbol{x}) = \prod_{l=1}^n q_i(\boldsymbol{y}_l)$, *then* $\delta\{p_1, p_2\} = n\,\delta\{q_1, q_2\}$.

Proof. (i) From Definition 3.1, $\delta\{p_1, p_2\}$ is obviously symmetric. (ii) Moreover, $\delta\{p_1, p_2\} = \min\{\kappa\{p_1 \,|\, p_2\}, \kappa\{p_2 \,|\, p_1\}\}$; but $\kappa\{p_i \,|\, p_j\}$ is non-negative (use the inequality $\log w \leq w - 1$ with $w = p_i/p_j$, multiply by p_j and integrate), and vanishes if (and only if) $p_i(\boldsymbol{x}) = p_j(\boldsymbol{x})$ almost everywhere. (iii) If $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ have strictly nested supports, one of the two directed divergences will not be finite, and their intrinsic discrepancy simply reduces to the other directed divergence. (iv) The new densities are $q_i(\boldsymbol{x}) = p_i(\boldsymbol{x})/|J|$, where J is the jacobian of the transformation; hence,

$$\kappa\{q_i \,|\, q_j\} = \int_{\boldsymbol{\mathcal{Z}}} p_j(\boldsymbol{x})\, |J| \log \frac{p_j(\boldsymbol{x})\, |J|}{p_i(\boldsymbol{x})\, |J|}\, \mathrm{d}\boldsymbol{z} = \int_{\boldsymbol{\mathcal{X}}} p_j(\boldsymbol{x}) \log \frac{p_j(\boldsymbol{x})}{p_i(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x}$$

which is $\kappa\{p_i \,|\, p_j\}$. (v) Under independence, $p_i(\boldsymbol{x}) = \prod_{j=1}^n q_i(\boldsymbol{y}_j)$; thus

$$\begin{aligned}
\kappa\{p_i \,|\, p_j\} &= \int_{\boldsymbol{\mathcal{Y}}^n} \prod_{l=1}^n q_j(\boldsymbol{y}_l) \log \frac{\prod_{l=1}^n q_j(\boldsymbol{y}_l)}{\prod_{l=1}^n q_i(\boldsymbol{y}_l)}\, \mathrm{d}\boldsymbol{y}_1 \ldots \mathrm{d}\boldsymbol{y}_l \\
&= n \int_{\boldsymbol{\mathcal{Y}}} q_j(\boldsymbol{y}) \log \frac{q_j(\boldsymbol{y})}{q_i(\boldsymbol{y})}\, \mathrm{d}\boldsymbol{y} = n\, \kappa\{q_i \,|\, q_j\}
\end{aligned}$$

and the result follows from Definition 3.1. □

The statistically important additive property is essentially unique to logarithmic discrepancies; it is basically a consequence of two facts (i) the joint density of independent random quantities is the product of their marginals, and (ii) the logarithm is the only analytic function which transforms products into sums.

The intrinsic discrepancy may be used to define a new type of convergence for probability distributions which finds many applications in both probability theory and Bayesian inference.

Definition 3.2 (Intrinsic convergence). *A sequence of probability distributions specified by their density functions* $\{p_i(\boldsymbol{x})\}_{i=1}^\infty$ *is said to* converge intrinsically *to a probability distribution with density* $p(\boldsymbol{x})$ *whenever the sequence of their intrinsic discrepancies* $\{\delta(p_i,\, p)\}_{i=1}^\infty$ *converges to zero.*

Example 3.1 (Poisson approximation to a Binomial distribution). *The intrinsic discrepancy between a Binomial distribution with probability function* $\mathrm{Bi}(r \,|\, n, \theta)$ *and its Poisson approximation* $\mathrm{Po}(r \,|\, n\,\theta)$*, is*

$$\delta\{\mathrm{Bi}, \mathrm{Po} \,|\, n, \theta\} = \sum_{r=0}^n \mathrm{Bi}(r \,|\, n,\, \theta) \log \frac{\mathrm{Bi}(r \,|\, n,\, \theta)}{\mathrm{Po}(r \,|\, n\,\theta)}\ ,$$

since the second sum in Definition 3.1 diverges. It may easily be verified that $\lim_{n\to\infty} \delta\{\mathrm{Bi}, \mathrm{Po}\,|\, n, \lambda/n\} = 0$ *and* $\lim_{\theta\to 0} \delta\{\mathrm{Bi}, \mathrm{Po}\,|\, \lambda/\theta,\, \theta\} = 0$*; thus, the sequences of Binomials* $\mathrm{Bi}(r\,|\, n, \lambda/n)$ *and* $\mathrm{Bi}(r\,|\, \lambda/\theta_i,\, \theta_i)$ *both intrinsically converge to a Poisson* $\mathrm{Po}(r\,|\,\lambda)$ *when* $n \to \infty$ *and* $\theta_i \to 0$*, respectively. Notice however that in the approximation a Binomial* $\mathrm{Bi}(r\,|\, n, \theta)$ *by a Poisson* $\mathrm{Po}(r\,|\, n\,\theta)$ *the rôles of* n *and* θ *are very far from similar; the crucial condition for the approximation to work is that the value of* θ *must be small, while the value of* n *is largely irrelevant. Indeed, as shown in Figure 4,* $\lim_{\theta\to 0} \delta\{\mathrm{Bi}, \mathrm{Po}\,|\, n,\, \theta\} = 0$*, for all* $n > 0$*, so arbitrarily good approximations are possible with any* n*, provided* θ *is sufficiently small. However,* $\lim_{n\to\infty} \delta\{\mathrm{Bi}, \mathrm{Po}\,|\, n, \theta\} = \frac{1}{2}[-\theta - \log(1-\theta)]$ *for all* $\theta > 0$*; thus, for fixed* θ*, the quality of the approximation cannot improve over a certain limit, no matter how large* n *might be.*

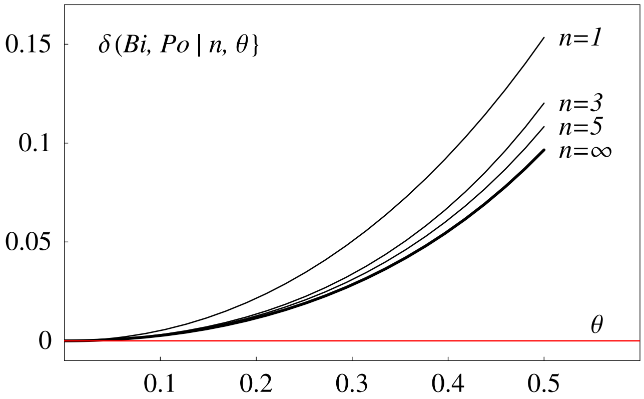

Figure 4: *Intrinsic discrepancy* $\delta\{\mathrm{Bi}, \mathrm{Po}\,|\, n, \theta\}$ *between a Binomial* $\mathrm{Bi}(r\,|\, n, \theta)$ *and a Poisson* $\mathrm{Po}(r\,|\, n\theta)$ *as a function of* θ*, for* $n = 1, 3, 5$ *and* ∞*.*

Definition 3.3 (Intrinsic discrepancy loss). *For any given parametric model* $\mathcal{M} = \{p(\boldsymbol{x}\,|\,\boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$*, the intrinsic discrepancy loss associated to the use of* $\boldsymbol{\omega}_0$ *as a proxy for* $\boldsymbol{\omega}$ *is the intrinsic discrepancy*

$$\delta_{\boldsymbol{x}}\{\boldsymbol{\omega}_0, \boldsymbol{\omega}\} = \delta\{p_{\boldsymbol{x}\,|\,\boldsymbol{w}_0},\, p_{\boldsymbol{x}\,|\,\boldsymbol{w}}\}$$

between the models identified by $\boldsymbol{\omega}_0$ *and* $\boldsymbol{\omega}$*. More generally, if* $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{\lambda})$*, so that the model is* $\mathcal{M} \equiv \{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\theta} \in \boldsymbol{\Theta},\, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$*, the intrinsic discrepancy loss associated to the use of* $\boldsymbol{\theta}_0$ *as a proxy for* $\boldsymbol{\theta}$ *is the intrinsic discrepancy*

$$\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \delta\{p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0},\, p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}}\}$$

between the assumed model $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})$, and its closest element in the family $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\lambda_0}), \boldsymbol{\lambda_0} \in \boldsymbol{\Lambda}\}$.

Example 3.2 (Intrinsic discrepancy loss in a Binomial model). *The intrinsic discrepancy loss $\delta_r\{\theta_0, \theta \mid n\}$ associated to the use of θ_0 as a proxy for θ with Binomial $\mathrm{Bi}(r \mid n, \theta)$ data is*

$$\begin{aligned} \delta_r\{\theta_0, \theta \mid n\} &= n\, \delta_x\{\theta_0, \theta\}, \\ \delta_x\{\theta_0, \theta\} &= \min[\, \kappa\{\theta_0 \mid \theta\},\, \kappa\{\theta \mid \theta_0\}\,] \\ \kappa(\theta_i \mid \theta_j) &= \theta_j \log[\theta_j/\theta_i] + (1-\theta_j)\log[(1-\theta_j)/(1-\theta_i)], \end{aligned} \tag{3.5}$$

where $\delta_x\{\theta_0, \theta\}$ is the intrinsic discrepancy between Bernoulli random variables with parameters θ_0 and θ. The intrinsic loss function $\delta_x\{\theta_0, \theta\}$ is represented in Figure 5.

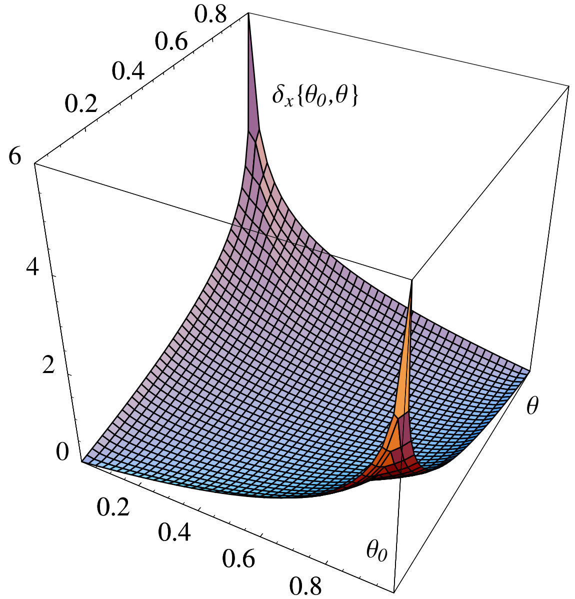

Figure 5: Intrinsic discrepancy loss $\delta_x\{\theta_0, \theta\}$ from using θ_0 as a proxy for θ in a binomial setting.

The *intrinsic discrepancy loss*, was introduced by Bernardo and Rueda (2002) in the context of hypothesis testing. It is an intrinsic loss function (Definition 2.2) and, hence, it is invariant under reparametrization. Moreover, as one would surely require, (i) the intrinsic discrepancy between two elements of a parametric family of distributions is also invariant under marginalization to the model induced by the sampling distribution of any sufficient statistic, and (ii) the intrinsic discrepancy loss is additive for conditionally independent observations. More precisely,

Theorem 3.2 (Properties of the intrinsic discrepancy loss). *Consider model $\mathcal{M} \equiv \{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ and let $\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ be the loss associated to the use of $\boldsymbol{\theta}_0$ as a proxy for $\boldsymbol{\theta}$.*

(i). Consistent marginalization: If $\boldsymbol{t}(\boldsymbol{x}) \in \boldsymbol{\mathcal{T}}$ is sufficient for $\mathcal{M}$, then $\delta_{\boldsymbol{t}}\{\boldsymbol{\theta}_0,\ (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$. In particular, $\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ is invariant under one-to-one transformations of $\boldsymbol{x}$.

(ii). Additivity: If $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ and the $\boldsymbol{y}_j$'s are independent given $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, then $\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \sum_{j=1}^{n} \delta_{\boldsymbol{y}_j}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$. If they are also identically distributed, then $\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = n\, \delta_{\boldsymbol{y}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$.

Proof. (i) If $\boldsymbol{t}(\boldsymbol{x}) \in \boldsymbol{\mathcal{T}}$ is sufficient for $\mathcal{M}$ and $\boldsymbol{s}(\boldsymbol{x})$ is an ancillary statistic, so that, in terms of $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{\lambda})$, $p(\boldsymbol{x}\,|\,\boldsymbol{\omega}) = p(\boldsymbol{t}\,|\,\boldsymbol{\omega})\, p(\boldsymbol{s}\,|\,\boldsymbol{t})$, the required directed divergences $\kappa\{p(\boldsymbol{x}\,|\,\boldsymbol{\omega}_i)\,|\,p(\boldsymbol{x}\,|\,\boldsymbol{\omega}_j)\}$ may be written as

$$\int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{t}\,|\,\boldsymbol{\omega}_j)\, p(\boldsymbol{s}\,|\,\boldsymbol{t}) \log \frac{p(\boldsymbol{t}\,|\,\boldsymbol{\omega}_j)\, p(\boldsymbol{s}\,|\,\boldsymbol{t})}{p(\boldsymbol{t}\,|\,\boldsymbol{\omega}_i)\, p(\boldsymbol{s}\,|\,\boldsymbol{t})}\, \mathrm{d}\boldsymbol{x} = \int_{\boldsymbol{\mathcal{T}}} p(\boldsymbol{t}\,|\,\boldsymbol{\omega}_j) \log \frac{p(\boldsymbol{t}\,|\,\boldsymbol{\omega}_j)}{p(\boldsymbol{t}\,|\,\boldsymbol{\omega}_i)}\, \mathrm{d}\boldsymbol{t}.$$

It follows that the intrinsic discrepancy loss $\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ calculated from the full model is the same as the intrinsic discrepancy $\delta_{\boldsymbol{t}}\{\boldsymbol{\theta}_0,\ (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ calculated from the marginal model $\{p(\boldsymbol{t}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{t} \in \boldsymbol{\mathcal{T}},\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ induced by the sufficient statistic. (ii) Additivity is a direct consequence of the last statement in Theorem 3.1. □

Computation of intrinsic loss functions in well-behaved problems may be simplified by the use of the result below:

Theorem 3.3 (Computation of the intrinsic loss function). *Consider a model $\mathcal{M} \equiv \{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ such that the support of $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})$ is convex for all pairs $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Then*

$$\begin{aligned}
\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} &= \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \delta\{p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0},\, p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}}\} \\
&= \min\left\{ \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \kappa(\boldsymbol{\theta}, \boldsymbol{\lambda}\,|\,\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0),\ \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \kappa(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}) \right\}.
\end{aligned}$$

Proof. This follows from the fact that the directed divergences are convex functions which are bounded above zero; for details, see Juárez (2004). □

Example 3.3 (Intrinsic discrepancy loss in a normal model). *By Theorems 3.2 and 3.3, the intrinsic discrepancy loss* $\delta_{\boldsymbol{x}}\{\mu_0, (\mu, \sigma)\}$ *associated to the use of* μ_0 *as a proxy for* μ *with a random sample* $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ *of normal* $N(x\,|\,\mu, \sigma)$ *data is* $n\,\delta_x\{\mu_0, (\mu, \sigma)\}$*, where*

$$\delta_x\{\mu_0, (\mu, \sigma)\} = \min[\inf_{\sigma_0>0} \kappa\{\mu, \sigma\,|\,\mu_0, \sigma_0\},\ \inf_{\sigma_0>0} \kappa\{\mu_0, \sigma_0\,|\,\mu, \sigma\}].$$

If σ *is known, then the two directed divergences are equal; indeed, given* σ*,*

$$\kappa\{\mu_i\,|\,\mu_j\} = \int_{\mathbb{R}} \mathrm{N}(x\,|\,\mu_j, \sigma) \log \frac{\mathrm{N}(x\,|\,\mu_j, \sigma)}{\mathrm{N}(x\,|\,\mu_i, \sigma)}\, dx = \frac{1}{2}\frac{(\mu_i - \mu_j)^2}{\sigma^2}$$

and, therefore,

$$\delta_{\boldsymbol{x}}\{\mu_0, \mu\,|\,\sigma)\} = \frac{n}{2}\left[\frac{(\mu_0 - \mu)^2}{\sigma^2}\right] = \frac{1}{2}\left[\frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right]^2, \tag{3.6}$$

just one half the square of the standardized difference between μ_0 *and* μ*.*

If σ *is not known, using Theorem 3.3,*

$$\begin{aligned}
\inf_{\sigma_0>0} \kappa\{\mu_0, \sigma_0\,|\,\mu, \sigma\} &= \inf_{\sigma_0>0} \int_{\mathbb{R}} \mathrm{N}(x\,|\,\mu, \sigma) \log \frac{\mathrm{N}(x\,|\,\mu, \sigma)}{\mathrm{N}(x\,|\,\mu_0, \sigma_0)}\, dx \\
&= \frac{1}{2} \log\left[1 + \frac{(\mu - \mu_0)^2}{\sigma^2}\right] \qquad (3.7) \\
\inf_{\sigma_0>0} \kappa\{\mu, \sigma\,|\,\mu_0, \sigma_0\} &= \inf_{\sigma_0>0} \int_{\mathbb{R}} \mathrm{N}(x\,|\,\mu_0, \sigma_0) \log \frac{\mathrm{N}(x\,|\,\mu_0, \sigma_0)}{\mathrm{N}(x\,|\,\mu, \sigma)}\, dx \\
&= \frac{1}{2}\left[\frac{(\mu - \mu_0)^2}{\sigma^2}\right]. \qquad (3.8)
\end{aligned}$$

Since, for all $w > 0$*,* $w \geq \log(1+w)$*, this implies that the required minimum is achieved by (3.7) and, therefore,*

$$\delta_{\boldsymbol{x}}\{\mu_0, (\mu, \sigma)\} = \frac{n}{2} \log\left[1 + \frac{(\mu - \mu_0)^2}{\sigma^2}\right], \tag{3.9}$$

a one-to-one function $\delta(z, n) = (n/2)\log[1+z^2]$ *of the Mahalanobis distance* $z^2 = (\mu - \mu_0)^2/\sigma^2$ *between* $\mathrm{N}(x\,|\,\mu_0, \sigma)$ *and* $\mathrm{N}(x\,|\,\mu, \sigma)$*. This generalizes to a multivariate normal setting.*

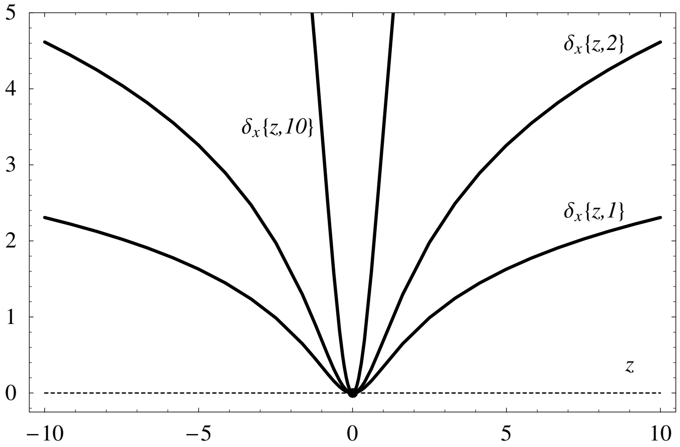

Figure 6: *Intrinsic discrepancy loss* $\delta_{\boldsymbol{x}}\{z, n\}$ *of using* μ_0 *as a proxy for* μ *given* n *normal* $\mathrm{N}(x \mid \mu, \sigma)$ *observations, as a function of* $z = |\mu_0 - \mu|/\sigma$, *for* $n = 1, 2, 10$.

The intrinsic discrepancy loss function $\delta_{\boldsymbol{x}}\{\mu_0, (\mu, \sigma)\}$ *is represented in Figure 6 as a function of the standardized distance* $z = (\mu_0 - \mu)/\sigma$ *between* μ_0 *and* μ, *for several values of* n. *Notice that for* $|z| \geq 1$, *the intrinsic discrepancy loss is concave, showing a very reasonable decreasing marginal loss, which is not present in conventional loss functions.*

4 Intrinsic credible regions

Lowest posterior loss credible regions (Definition 2.1) depend both on the loss function and on the prior distribution. It has been argued that, in scientific inference, loss functions should be invariant under reparametrization; this is always achieved by *intrinsic loss functions* (Definition 2.2), which measure the discrepancy between the models identified by the parameters, rather than the discrepancy between the parameters themselves. It has further been argued that intrinsic loss functions should be required to be symmetric and consistent with the use of sufficient statistics. The *intrinsic discrepancy loss* (Definition 3.3) meets these requirements, and has many additional attractive properties, notably its additivity under conditional independence. It may therefore be reasonable to propose the intrinsic discrepancy loss as an appropriate conventional loss for routine use in mathematical statistics.

On the other hand, as already mentioned in the introduction, scientific communication typically requires the use of some sort objective prior, one which captures, in a well-defined sense, the notion of the prior having a minimal effect, relative to the data in the final inference. This should be a conventional prior to be used when a default specification, having a claim to being non-influential in the sense described above, is required. In the long historical quest for these objective priors several requirements have emerged which may reasonably be requested as necessary properties of any proposed solution; this includes generality, consistency under reparametrization, consistency under marginalization, and consistent sampling properties. *Reference analysis*, introduced by Bernardo (1979b) and further developed Berger and Bernardo (1989, 1992a,b,c), appears to be the only available method to derive objective prior functions which satisfy all these desiderata. For an introduction to reference analysis, see Bernardo and Ramón (1998); for a recent review of reference analysis, see Bernardo (2005b).

The Bayes estimator which corresponds to the intrinsic discrepancy loss and the appropriate reference prior is the *intrinsic estimator*. Introduced by Bernardo and Juárez (2003), this is a completely general objective Bayesian estimator, which is invariant under reparametrization.

Definition 4.1 (Intrinsic estimate). *Consider data* $\boldsymbol{x}$ *which consist of one observation from* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$*, and let* $\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ *be the intrinsic discrepancy loss to be suffered if* $\boldsymbol{\theta}_0$ *were used as a proxy for* $\boldsymbol{\theta}$*. The* intrinsic estimate *of* $\boldsymbol{\theta}$

$$\boldsymbol{\theta}^*(\boldsymbol{x}) = \arg \min_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}} d(\boldsymbol{\theta}_i \,|\, \boldsymbol{x}),$$

is that parameter value which minimizes the reference posterior expected intrinsic loss $d(\boldsymbol{\theta}_i \,|\, \boldsymbol{x})$*, where*

$$\begin{aligned} d(\boldsymbol{\theta}_i \,|\, \boldsymbol{x}) &= \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_i,\ (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}, && (4.1) \\ \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x}) &\propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}), && (4.2) \end{aligned}$$

and $\pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$ *is the joint reference prior of* $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ *when* $\boldsymbol{\theta}$ *is the quantity of interest.*

Moving from point estimation to region estimation, *intrinsic credible regions* are defined as the lowest posterior loss credible regions which correspond to the use of the intrinsic discrepancy loss and the appropriate reference prior. As one would expect, the intrinsic estimate is contained in all intrinsic credible regions.

Definition 4.2 (Intrinsic credible region). *Consider data $\boldsymbol{x}$ which consist of one observation from $\mathcal{M} \equiv \{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\theta} \in \boldsymbol{\Theta},\, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, and let $\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ be the intrinsic discrepancy loss to be suffered if $\boldsymbol{\theta}_0$ were used as a proxy for $\boldsymbol{\theta}$. An* intrinsic q-credible region *is a subset $R_q^* = R_q^*(\boldsymbol{x}, \boldsymbol{\Theta}) \subset \boldsymbol{\Theta}$ of the parameter space $\boldsymbol{\Theta}$ such that,*

(i). $\int_{R_q^*} \pi(\boldsymbol{\theta}\,|\,\boldsymbol{x})\, d\boldsymbol{\theta} = q$,

(ii). $\forall \boldsymbol{\theta}_i \in R_q^*,\ \forall \boldsymbol{\theta}_j \notin R_q^*,\ d(\boldsymbol{\theta}_i\,|\,\boldsymbol{x}) \le d(\boldsymbol{\theta}_j\,|\,\boldsymbol{x})$,

where $d(\boldsymbol{\theta}_i\,|\,\boldsymbol{x})$, the reference intrinsic posterior expected loss from using $\boldsymbol{\theta}_i$ as a proxy for the value of the parameter, is given by Equation (4.1).

The analytical expression of the intrinsic discrepancy loss $\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0,\, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ is often complicated and, hence, exact computation of its posterior expectation, $d(\boldsymbol{\theta}_i\,|\,\boldsymbol{x})$ typically requires numerical integration. Although these days this is seldom a serious practical problem, it is both theoretically interesting and pragmatically useful to derive appropriate asymptotic approximations. Attention to approximations will be limited here to one-dimensional regular models, but the results may be extended to both non-regular and multiparameter problems.

Let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, $x_j \in \mathcal{X}$, consist of a random sample of size n from a distribution $p(x\,|\,\theta)$ with one continuous parameter $\theta \in \Theta \subset \mathbb{R}$. Under appropriate regularity conditions, there exists a unique maximum likelihood estimator $\hat{\theta}_n = \hat{\theta}_n(\boldsymbol{x})$ whose sampling distribution is asymptotically normal with mean θ and variance $i^{-1}(\theta)/n$, where $i(\theta)$ is Fisher's information function,

$$i(\theta) = -\int_{\mathcal{X}} p(x\,|\,\theta)\, \frac{\partial^2}{\partial\theta^2} \log p(x\,|\,\theta)\, \mathrm{d}x. \tag{4.3}$$

Moreover, the function defined by the indefinite integral

$$\phi(\theta) = \int \sqrt{i(\theta)}\, \mathrm{d}\theta \tag{4.4}$$

provides a variance stabilizing transformation. Indeed, it is easily verified that, under the assumed conditions, the maximum likelihood estimate

$\hat{\phi}_n = \hat{\phi}_n(\boldsymbol{x})$ is asymptotically normal with mean $\phi = \phi(\theta)$ and variance $1/n$, and that the approximate marginal model

$$p(\hat{\phi}_n \,|\, \phi) = \mathrm{N}(\hat{\phi}_n \,|\, \phi, 1/\sqrt{n}), \quad \phi \in \Phi = \phi(\Theta),$$

is asymptotically equivalent to the original model $p(\boldsymbol{x} \,|\, \theta) = \prod_{j=1}^{n} p(x_j \,|\, \theta)$. It follows that ϕ asymptotically behaves as location parameter and, hence, the reference prior for ϕ is the uniform prior $\pi(\phi) = 1$. All this suggests that $\phi(\theta)$ is, in a sense, a fairly natural parametrization for the model.

More generally, if $\tilde{\theta}_n = \tilde{\theta}_n(\boldsymbol{x})$ is an asymptotically sufficient, consistent estimator of θ whose asymptotic sampling distribution is $p(\tilde{\theta}_n \,|\, \theta)$, the reference prior for θ is (Bernardo and Smith, 1994, Section 5.4)

$$\pi(\theta) = p(\tilde{\theta}_n \,|\, \theta)\Big|_{\tilde{\theta}_n = \theta} \tag{4.5}$$

and, therefore, the reference prior of the monotone transformation defined by the indefinite integral $\phi(\theta) = \int \pi(\theta)\, \mathrm{d}\theta$ is $\pi(\phi) = \pi(\theta)/|\partial\phi/\partial\theta| = 1$, a uniform prior.

The *reference parametrization* of a probability model is defined as that for which the reference prior is uniform:

Definition 4.3 (Reference parametrization). *Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from $\mathcal{M} = \{p(x \,|\, \theta),\, x \in \mathcal{X},\, \theta \in \Theta\}$ and let $\tilde{\theta}_n = \tilde{\theta}_n(\boldsymbol{x})$ be an asymptotically sufficient, consistent estimator of θ whose asymptotic sampling distribution is $p(\tilde{\theta}_n \,|\, \theta)$.* A reference parametrization *for model $\mathcal{M}$ is then defined by the indefinite integral*

$$\phi(\theta) = \int \pi(\theta)\, d\theta, \quad \textit{where} \quad \pi(\theta) = p(\tilde{\theta}_n \,|\, \theta)\Big|_{\tilde{\theta}_n = \theta}. \tag{4.6}$$

When the sample space $\mathcal{X}$ does not depend on θ and the likelihood function $p(\boldsymbol{x} \,|\, \theta)$ is twice differentiable as a function of θ, the sampling distribution of maximum-likelihood estimator $\hat{\theta}$ is often asymptotically normal with variance $i^{-1}(\theta)/n$, where $i(\theta)$ is Fisher's information function given by Equation (4.3); see, *e.g.*, Schervish (1995, Section 7.3.5) for precise conditions. In this case, the reference parametrization is given by Equation (4.4), and this may be used to obtain analytical approximations. More generally, if a model has an asymptotically sufficient, consistent estimator of θ whose sampling distribution is asymptotically normal, a reference parametrization may be used to obtain a simple asymptotic approximation to its intrinsic discrepancy loss $\delta_{\boldsymbol{x}}(\theta_0, \theta)$, and to the corresponding reference posterior expectation $d(\theta_0 \,|\, \boldsymbol{x})$. This provides analytical asymptotic approximations to the required credible regions.

Theorem 4.1 (Asymptotic approximations). *Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from $\mathcal{M} = \{p(x \,|\, \theta), x \in \mathcal{X}, \theta \in \Theta\}$ and let $\tilde{\theta}_n = \tilde{\theta}_n(\boldsymbol{x})$ be an asymptotically sufficient, consistent estimator of θ whose sampling distribution is asymptotically normal $\mathrm{N}(\tilde{\theta}_n \,|\, \theta, s(\theta)/\sqrt{n})$. Then,*

(i). The reference prior for θ is $\pi(\theta) = s^{-1}(\theta)$, a reference parametrization is $\phi(\theta) = \int s^{-1}(\theta)\, d\theta$, the reference prior for ϕ is $\pi(\phi) = 1$, and the reference posterior of ϕ, in terms of the inverse function $\theta(\phi)$, is $\pi(\phi \,|\, \boldsymbol{x}) \propto p\{\boldsymbol{x} \,|\, \theta(\phi)\}\, |\partial\theta(\phi)/\partial\phi|$.

(ii). The intrinsic discrepancy loss is
$\delta_{\boldsymbol{x}}(\theta_0, \theta) = \frac{n}{2}\big[\,\phi(\theta_0) - \phi(\theta)\big]^2 + o(1)$.

(iii). The expected posterior loss is
$d(\theta_0 \,|\, \boldsymbol{x}) = \frac{n}{2}\big[\,\sigma^2_\phi(\boldsymbol{x}) + \{\mu_\phi(\boldsymbol{x}) - \phi(\theta_0)\}^2\,\big] + o(1)$,
where $\mu_\phi(\boldsymbol{x})$ and $\sigma^2_\phi(\boldsymbol{x})$ are, respectively, the mean and variance of the reference posterior distribution of ϕ, $\pi(\phi \,|\, \boldsymbol{x})$.

(iv). The intrinsic estimator of ϕ is $\mu_\phi(\boldsymbol{x}) + o(1)$, and the intrinsic estimator of θ is $\theta^*(\boldsymbol{x}) = \theta\{\mu_\phi(\boldsymbol{x})\} + o(1)$

(v). The intrinsic q-credible region of ϕ is the interval
$[\phi_{q0}(\boldsymbol{x}),\ \phi_{q1}(\boldsymbol{x})] = \mu_\phi(\boldsymbol{x}) \pm z_q\, \sigma_\phi(\boldsymbol{x}) + o(1)$,
where z_q is the $(q+1)/2$ normal quantile.

The intrinsic q-credible region of θ is the interval
$[\theta_{q0}(\boldsymbol{x}),\ \theta_{q1}(\boldsymbol{x})] = \theta\{\,[\phi_{q0}(\boldsymbol{x}),\ \phi_{q1}(\boldsymbol{x})]\,\} + o(1)$.

Proof. (i) is an immediate application of Equation (4.5), Definition 4.3 and standard probability calculus. (ii) Under the assumed conditions, the sampling distribution of $\tilde{\phi}_n(\boldsymbol{x})$ will be, for sufficiently large n, approximately normal $\mathrm{N}(\tilde{\phi}_n \,|\, \phi, 1/\sqrt{n})$. Since the intrinsic discrepancy loss is invariant under marginalization (Theorem 3.2), $\delta_{\boldsymbol{x}}(\theta_0, \theta) = \delta_{\tilde{\phi}_n}(\phi_0, \phi)$ and, using Equation (3.6) of Example 3.3,

$$\delta_{\tilde{\phi}_n}(\phi_0, \phi) \approx \delta_{\boldsymbol{x}}\{\mathrm{N}(\tilde{\phi}_n \,|\, \phi_0, 1/\sqrt{n}),\ \mathrm{N}(\tilde{\phi}_n \,|\, \phi, 1/\sqrt{n})\} = \frac{n}{2}\,(\phi_0 - \phi)^2.$$

(iii) Using the invariance of the intrinsic discrepancy loss under reparametrization,

$$
\begin{aligned}
d(\theta_0\,|\,\boldsymbol{x}) &= \int_\Theta \delta_{\boldsymbol{x}}(\theta_0,\theta)\,\pi(\theta\,|\,\boldsymbol{x})\,\mathrm{d}\theta = \int_\Phi \delta_{\boldsymbol{x}}(\phi_0,\phi)\,\pi(\phi\,|\,\boldsymbol{x})\,\mathrm{d}\phi \\
&\approx \int_\Phi \frac{n}{2}\,(\phi_0-\phi)^2\;\;\pi(\phi\,|\,\boldsymbol{x})\,\mathrm{d}\phi \\
&= \frac{n}{2}\big[\mathrm{E}(\phi-\mu_\phi)^2+(\mu_\phi-\phi_0)^2\big] = \frac{n}{2}\big[\sigma^2_\phi+(\mu_\phi-\phi_0)^2\big],
\end{aligned}
$$

where $\phi_0=\phi(\theta_0)$. (iv) As a function of ϕ_0, $d(\phi_0\,|\,\boldsymbol{x})$ is minimised when $\phi_0=\mu_\phi$ and, hence this provides the intrinsic estimate of ϕ; by invariance, the intrinsic estimate of θ is simply $\theta(\mu_\phi)$, where $\theta(\phi)$ is the inverse function of $\phi(\theta)$. (v) Since the expected intrinsic loss $d(\phi_0\,|\,\boldsymbol{x})$ is symmetric around μ_ϕ, all lowest posterior loss credible regions will be symmetric around μ_ϕ Hence, the intrinsic q-credible interval for ϕ will be of the form $R^*_q(\boldsymbol{x},\Phi)=\mu_\phi\pm z_q\,\sigma_\phi$, with z_q chosen such that

$$
\int_{\mu_\phi-z_q\,\sigma_\phi}^{\mu_\phi+z_q\,\sigma_\phi}\pi(\phi\,|\,\boldsymbol{x})\,\mathrm{d}\phi = q.
$$

Moreover, since $\tilde\phi_n(\boldsymbol{x})$ is asymptotically sufficient, and its sampling distribution is asymptotically normal, the reference posterior distribution of ϕ will also be asymptotically normal and, therefore, z_q will approximately be the $(q+1)/2$ quantile of the standard normal distribution. By invariance, the intrinsic q-credible interval for θ will simply be given by inverse image of q-credible interval for ϕ, $R^*_q(\boldsymbol{x},\Theta)=\theta\{R^*_q(\boldsymbol{x},\Phi)\}$. □

The posterior moments $\mu_\phi(\boldsymbol{x})$ and $\sigma^2_\phi(\boldsymbol{x})$ of the reference parameter required in Theorem 4.1 may often be obtained analytically. If this is not the case, the delta method may be used to derive $\mu_\phi(\boldsymbol{x})$ and $\sigma^2_\phi(\boldsymbol{x})$ in terms of the (typically easier to obtain) reference posterior mean $\mu_\theta(\boldsymbol{x})$ and reference posterior variance $\sigma^2_\theta(\boldsymbol{x})$ of the original parameter θ:

$$
\begin{aligned}
\mu_\phi(\boldsymbol{x}) &\approx \phi\{\mu_\theta(\boldsymbol{x})\}+\tfrac{1}{2}\,\sigma^2_\theta(\boldsymbol{x})\,\phi''\{\mu_\theta(\boldsymbol{x})\} && (4.7)\\
\sigma^2_\phi(\boldsymbol{x}) &\approx \sigma^2_\theta(\boldsymbol{x})\;[\phi'\{\mu_\theta(\boldsymbol{x})\}]^2 && (4.8)
\end{aligned}
$$

(see *e.g.*, Schervish (1995, Section 7.1.3) for precise conditions). The delta method yields a particularly simple approximation for the posterior variance of the reference parameter. Indeed, $\sigma^2_\theta\approx s(\tilde\theta_n)/n$ and $\phi^{'}(\theta)=s(\theta)$;

hence, using Equation (4.8), $\sigma_\phi^2 \approx 1/n$. This provides much simpler (but less precise) approximations than those in Theorem 4.1.

Corollary 4.1. *Under the conditions of Theorem 4.1, simpler (less precise) approximations are given by:*

(i). $d(\theta_0 \,|\, \boldsymbol{x}) \approx \frac{1}{2} + \frac{n}{2}\big[\,\mu_\phi(\boldsymbol{x}) - \phi(\theta_0)\big]^2$

(ii). $\theta^* \approx \theta\{\mu_\phi(\boldsymbol{x})\}$

(iii). $R_q^*(\boldsymbol{x}, \Phi) \approx \mu_\phi(\boldsymbol{x}) \pm z_q/\sqrt{n}, \quad R_q^*(\boldsymbol{x}, \Theta) = \theta\{R_q^*(\boldsymbol{x}, \Phi)\}$, *where* z_q *is the* $(q+1)/2$ *quantile of the standard normal distribution.*

(iv). If $\mu_\phi(\boldsymbol{x})$ *is not analytically available, it may be approximated in terms of the first reference posterior moments of* θ, $\mu_\theta(\boldsymbol{x})$ *and* $\sigma_\theta(\boldsymbol{x})$, *by* $\mu_\phi(\boldsymbol{x}) \approx \phi\{\mu_\theta(\boldsymbol{x})\} + \frac{1}{2}\,\sigma_\theta^2(\boldsymbol{x})\,\phi''\{\mu_\theta(\boldsymbol{x})\}$

As illustrated below, Corollary 4.1 may actually provide reasonable approximations even with rather small sample sizes.

Example 4.1 (Credible intervals for a binomial parameter (continued)). *Consider again the problem considered in Examples 1.1, 2.1 and 3.2, and use the corresponding intrinsic discrepancy loss (Equation 3.5). The reference prior is* $\pi(\theta) = \theta^{-1/2}(1-\theta)^{-1/2}$, *and the reference posterior is* $\pi(\theta \,|\, r, n) = \mathrm{Be}(\theta \,|\, r + \frac{1}{2}, n - r + \frac{1}{2})$. *The reference posterior expected loss from using* θ_0 *rather than* θ *will be*

$$d\{\theta_0 \,|\, r, n\} = n \int_0^1 \delta_x\{\theta_0, \theta\}\, \mathrm{Be}(\theta \,|\, r + \tfrac{1}{2}, n - r + \tfrac{1}{2})\, d\theta.$$

Simple numerical algorithms may be used to obtain the intrinsic estimate, namely the value of θ_0 *which minimizes* $d\{\theta_0 \,|\, r, n\}$, *and intrinsic credible intervals, that is the lowest posterior loss intervals with the required posterior probability.*

The function $d\{\theta_0 \,|\, r, n\}$ *is represented in the upper panel of Figure 7 for the case* $r = 0$ *and* $n = 10$ *discussed before. This is minimized by* $\theta^* = 0.031$, *which is therefore the intrinsic estimate; the result may be compared with the maximum-likelihood estimate* $\hat{\theta} = 0$, *utterly useless in this case. Similarly, for* $r = 1$ *and* $n = 10$ *the intrinsic estimate is found to be* $\theta^* = 0.123$; *by invariance, the intrinsic estimate of any one to one function* $\psi = \psi(\theta)$ *is simply* $\psi(\theta^*)$; *thus the intrinsic estimate of, say* θ^2, *is*

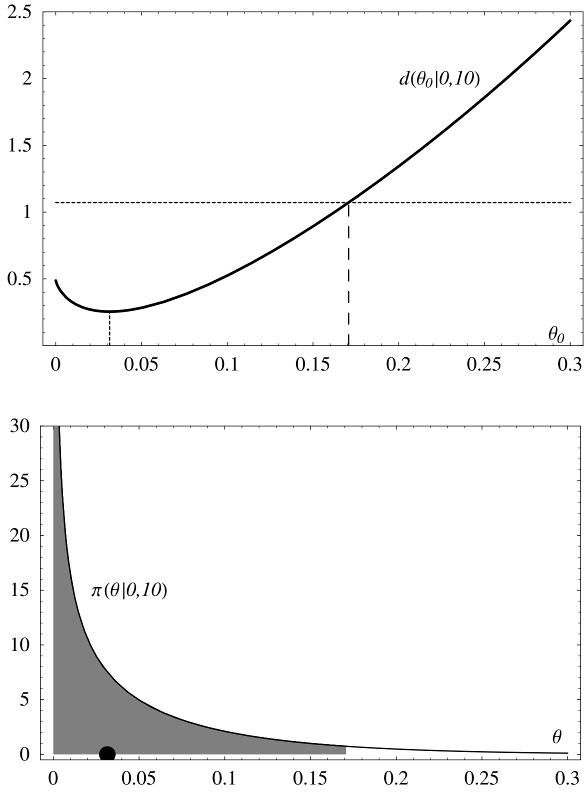

Figure 7: Reference posterior density, intrinsic estimate and intrinsic 0.95*-credible region for a binomial parameter, given* $n = 10$ *and* $r = 0$.

$\theta^{*2} = 0.015$*; this may be compared with the corresponding unbiased estimate of* θ^2*, which is* $\{r(r-1)\}/\{n(n-1)\}$ *and hence zero in this particular case, a rather obtuse estimation for the square of the proportion of something which has actually been observed one in ten times. For* $r = 2$*,* $n = 10$ *the intrinsic estimate is* $\theta^* = 0.218$*, somewhere between the posterior median* (0.210) *and the the posterior mean* (0.227)*.*

Intrinsic credible regions are also easily found numerically. Thus, for $r = 0$ *and* $n = 10$*, the* 0.95 *intrinsic credible interval for* θ *is* $(0, 0.170)$ *(shaded region in Figure 7); in this case, this is also the* HPD *interval. For* $r = 2$ *and* $n = 10$ *the* 0.95 *intrinsic credible interval is* $(0.032, 0.474)$*, very close to* $(0.037, 0.482)$*, the* LPL 0.95*-credible interval which corresponds to the* $\mathcal{L}_1$ *loss (see Example 2.1).*

Since the reference prior for this problem is $\pi(\theta) = \theta^{-1/2}(1-\theta)^{-1/2}$*, a reference parameter is* $\phi(\theta) = \int \pi(\theta)\, d\theta = 2\,\mathrm{arcsin}\sqrt{\theta}$*, with inverse function* $\theta(\phi) = \sin^2(\phi/2)$*. Changing variables, the reference posterior of the reference parameter* ϕ *is Equation* (1.4)*, a distribution whose first moments do not have a simple analytical expression. The use of Corollary* 4.1 *with the exact reference posterior moments of* θ*,* $\mu_\theta = (r+1/2)/(n+1)$ *and* $\sigma^2_\theta = \mu_\theta(1-\mu_\theta)/(n+2)$ *leads, with* $\phi^{'}(\theta) = \theta^{-1/2}(1-\theta)^{-1/2}$ *and* $\phi^{''}(\theta) = \frac{1}{2}\,(2\theta-1)\,\theta^{-3/2}(1-\theta)^{-3/2}$*, to simple analytical approximations for the intrinsic estimates and the intrinsic credible regions. In particular, with* $r=2$ *and* $n=10$*,* $\mu_\theta = 0.227$*,* $\sigma^2_\theta = 0.015$*, and the delta method yields* $\mu_\phi \approx 0.967$ *and* $\theta^* = \theta(\mu_\phi) \approx 0.216$*, quite close to the exact value* 0.218*. Moreover* $R^*_{0.95}(\Theta) \approx \theta\{\,(0.967 \pm 1.96 \times 1/\sqrt{10})\,\} = (0.030,\, 0.508)$*, close to its exact value* $(0.032,\, 0.474)$*. As one would expect the approximation is not that good in extreme cases, when either* $r=0$ *or* $r=n$*. For instance, with* $r=0$ *and* $n=10$*, Corollary* 4.1 *yields* $\theta^* \approx 0.028$ *and* $R^*_{0.95}(\Theta) \approx (0.020,\, 0.213)$*, compared with the exact values* 0.031 *and* $(0,\, 0.170)$ *respectively.*

Example 4.2 (Intrinsic credible interval for the normal mean). *Consider a random sample* $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ *from a normal distribution* $\mathrm{N}(x\,|\,\mu, \sigma)$*, and let* $\overline{x} = \Sigma_j x_j/n$*, and* $s^2 = \Sigma_j(x_j - \overline{x})^2/n$ *be the corresponding mean and variance. The reference prior when* μ *is the parameter of interest is* $\pi(\mu)\,\pi(\sigma\,|\,\mu) = \sigma^{-1}$*, and the corresponding joint reference posterior is*

$$\begin{aligned}\pi(\mu, \sigma\,|\,\boldsymbol{x}) \;&=\; \pi(\mu, \sigma\,|\,\boldsymbol{x}) = \mathrm{N}[\mu\,|\,\overline{x}, \sigma/\sqrt{n}]\;\mathrm{Ga}^{-1/2}[\sigma\,|\,\tfrac{1}{2}(n-1), \tfrac{1}{2}n\,s^2]\\ &\propto\; \sigma^{-(n+1)}\,\exp[-\frac{n}{\sigma^2}\{s^2 + (\overline{x}-\mu)^2\}]. \qquad (4.9)\end{aligned}$$

Thus, using Equation (3.9)*, the reference posterior expected intrinsic loss from using* μ_0 *as a proxy for* μ *is*

$$d(\mu_0\,|\,\boldsymbol{x}) = \int_\infty^\infty \int_0^\infty \frac{n}{2}\log\left[1 + \frac{(\mu-\mu_0)^2}{\sigma^2}\right]\,\pi(\mu, \sigma\,|\,\boldsymbol{x})\; d\mu\, d\sigma. \qquad (4.10)$$

As one could expect, and may directly be verified by appropriate change of variables in Equation (4.10)*, this is a symmetric function of* $\mu_0 - \overline{x}$*. It follows that q-credible regions for* μ *must be centred at* $\overline{x}$*. Moreover, the (marginal) reference posterior of* μ *is Student* $\mathrm{St}(\mu|\overline{x}, s/\sqrt{n-1}, n-1)$*. Consequently, the intrinsic q-credible interval for the normal mean* μ *is*

$$R^*_q(\boldsymbol{x}, \mathbb{R}) = \overline{x} \pm t_q(n-1)\;s/\sqrt{n-1}$$

where $t_q(n-1)$ *is the* $(q+1)/2$ *quantile of a standard Student distribution with* $n-1$ *degrees of freedom. As one could expect in this example, this is also the* HPD *interval, and the frequentist confidence interval of level* $1-q$.

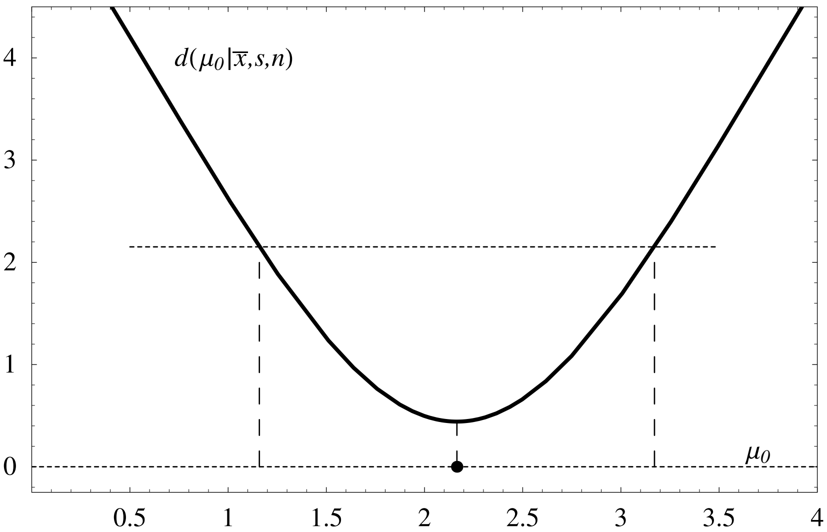

Figure 8: *Expected intrinsic loss from using* μ_i *as a proxy for a normal mean, given* $n = 10$ *observations, with* $\overline{x} = 2.165$ *and* $s = 1.334$.

Figure 8 *describes the behaviour of* $d(\mu_0 \,|\, \boldsymbol{x})$ *given* $n = 10$ *observations, simulated from* $\mathrm{N}(x \,|\, 2, 1)$ *which yielded* $\overline{x} = 2.165$ *and* $s = 1.334$*. The intrinsic estimate is obviously* $\overline{x}$ *(marked with a solid dot) and the* 0.95 *intrinsic credible interval consists of values* (1.159, 3.171) *whose intrinsic posterior expected loss is smaller than* 2.151.

5 Frequentist coverage

As Example 4.2 illustrates, the frequentist coverage probabilities of the q-credible regions which may be derived from objective posterior distributions are sometimes identical to their posterior probabilities. This *exact* numerical agreement is however the exception, not the norm. Nevertheless, for *large* sample sizes, Bayesian credible intervals are *always approximate* confidence intervals. Although this is an asymptotic property, it has been found that, even for moderate samples, the frequentist coverage of *reference* q-credible regions, *i.e.*, credible regions based of the reference posterior distribution, is usually very close to q. This means that, in many problems, reference q-credible regions are also *approximate* frequentist confidence intervals with significance level $1-q$; thus, under repeated sampling from the same model, the proportion of reference q-credible regions containing the true value of the parameter will be approximately q.

More precisely, let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consist of n independent observations from the one parameter model $\mathcal{M} = \{p(x \,|\, \theta), x \in \mathcal{X}, \theta \in \Theta\}$, and let $\theta_q(\boldsymbol{x},\, p_\theta)$ denote the q-quantile of the posterior $p(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\, p(\theta)$ which corresponds to the prior $p(\theta)$; thus,

$$\Pr\Big[\theta \leq \theta_q(\boldsymbol{x},\, p_\theta) \,|\, \boldsymbol{x}\Big] = \int_{\{\theta \leq \theta_q(\boldsymbol{x},\, p_\theta)\}} p(\theta \,|\, \boldsymbol{x})\, \mathrm{d}\theta = q.$$

and, for any fixed data $\boldsymbol{x}$, $R_q(\boldsymbol{x}) = \{\theta;\ \theta \leq \theta_q(\boldsymbol{x},\, p_\theta)\}$ is a left q-credible region for θ. For fixed θ, consider now $R_q(\boldsymbol{x})$ as a function of $\boldsymbol{x}$. Standard asymptotic theory may be used to establish that, for any sufficiently regular pair $\{p_\theta,\, \mathcal{M}\}$ of prior p_θ and model $\mathcal{M}$, the *coverage* probability of $R_q(\boldsymbol{x})$ converges to q as $n \to \infty$. Specifically, for all sufficiently regular priors,

$$\Pr\Big[\theta_q(\boldsymbol{x}, p_\theta) \geq \theta \,|\, \theta\Big] = \int_{\{\boldsymbol{x};\ \theta_q(\boldsymbol{x}, p_\theta) \geq \theta\}} p(\boldsymbol{x} \,|\, \theta)\, \mathrm{d}\boldsymbol{x} = q + O(n^{-1/2}).$$

In a pioneering paper, Welch and Peers (1963) established that in the case of the one-parameter regular models for continuous data Jeffreys prior, which in this case is also the reference prior, $\pi(\theta) \propto i(\theta)^{1/2}$, is the only prior which further satisfies

$$\Pr\Big[\theta_q(\boldsymbol{x}, \pi_\theta) \geq \theta \,|\, \theta\Big] = q + O(n^{-1});$$

Hartigan (1966) later showed that the coverage probabilities of one-dimensional *two-sided* Bayesian posterior credible intervals satisfy this type of approximation to $O(n^{-1})$ for *all* sufficiently regular prior functions. Moreover, Hartigan (1983, p. 79) showed that the result of Welch and Peers (1963) on one-sided posterior credible intervals may be extended to one-parameter models for discrete data by using appropriate continuity corrections.

This all means that reference priors are often *probability matching* priors, that is, priors for which the coverage probabilities of posterior credible intervals are *asymptotically* closer to their posterior probabilities than those derived from any other prior; see Datta and Sweeting (2005) for a recent review on this topic.

Although the results described above only justify an *asymptotic* approximate frequentist interpretation of reference credible regions, the coverage probabilities of reference q-credible regions derived from *relatively small samples* are found to be relatively close to their posterior probability q. This is now illustrated within the binomial parameter problem.

Example 5.1 (Frequentist coverage of binomial credible regions). *Consider again the intrinsic credible intervals for a binomial parameter of Example 4.1. The frequentist coverage of the intrinsic q-credible region $R_q^*(r, n, \Theta)$ there defined is*

$$\operatorname{Cov}\{R_q^* \,|\, \theta, n\} = \Pr[\theta \in R_q^*(r, n, \Theta) \,|\, \theta, n\,] = \sum_{\{r;\; \theta \in R_q^*\}} \tbinom{n}{r} \theta^r (1-\theta)^{n-r}.$$

Since r is discrete, this cannot be a continuous function of θ. Indeed the frequentist coverage $\operatorname{Cov}\{R_q^* \,|\, \theta, n\}$ *of the q-intrinsic credible region is bound to oscillate around its reference posterior probability q. It may be argued however that Bayesian reference posterior q-credible regions possibly provide the best available solution for this particular frequentist problem.*

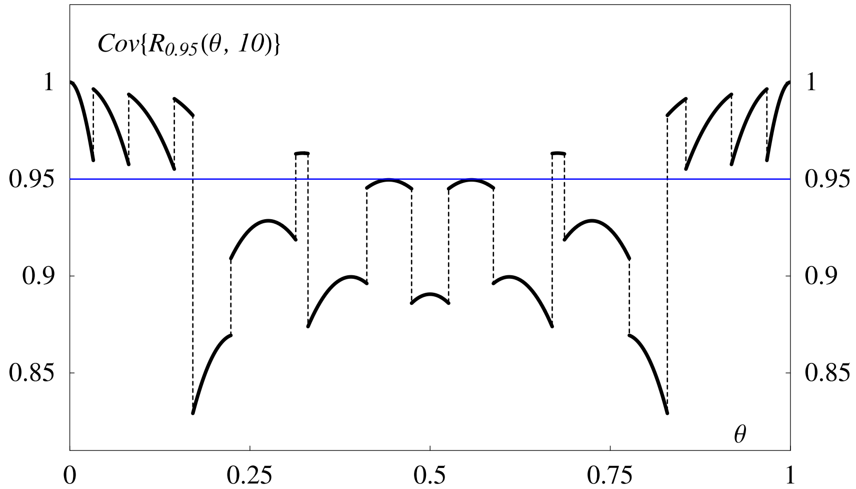

Figure 9: *Frequentist coverage* $\operatorname{Cov}\{R_q^* \,|\, \theta, n\}$ *of binomial* 0.95*-intrinsic credible regions with* $n = 10$.

For a numerical illustration consider again the case $n = 10$, so that $r \in \{0, 1, \ldots, 10\}$. Since there are only 11 different possible values of r,

Table 1: Intrinsic estimates and intrinsic 0.95*-credible intervals for the parameter* θ *of a binomial distribution* $\mathrm{Bi}(r \,|\, n, \theta)$*, with* $n = 10$.

r	$\theta^*(r, 10)$	$R_q^*(r, n, \Theta)$
0	0.032	(0.000, 0.171)
1	0.124	(0.000, 0.331)
2	0.218	(0.033, 0.474)
3	0.314	(0.082, 0.588)
4	0.408	(0.145, 0.686)
5	0.500	(0.224, 0.776)
6	0.592	(0.314, 0.855)
7	0.686	(0.412, 0.918)
8	0.782	(0.526, 0.967)
9	0.876	(0.669, 1.000)
10	0.968	(0.829, 1.000)

there are only 11 *distinct intrinsic* 0.95*-credible intervals; those are listed in Table 1. If the true value of* θ *were, say,* 0.25*, it would be contained in the intrinsic credible region* $R_q^*(r, n, \Theta)$ *if, and only if,* $r \in \{1, 2, 3, 4, 5\}$*, and this would happen with probability*

$$\mathrm{Cov}\{R_{0.95}^* \,|\, \theta = 0.25, n = 10\} = \sum_{r=1}^{5} \mathrm{Bi}(r \,|\, \theta = 0.25,\ n = 10) = 0.934.$$

Figure 9 represents the frequentist coverage $\mathrm{Cov}\{R_q^* \,|\, \theta, n\}$ *as a function of* θ *for* $q = 0.95$ *and* $n = 10$*. It may be appreciated the proportion* $\mathrm{Cov}\{R_{0.95}^* \,|\, \theta, n\}$ *of intrinsic* 0.95*-credible intervals which may be expected to contain the true value of* θ *under repeated sampling oscillates rather wildly around its posterior probability* 0.95*, with discontinuities at the points which define the credible regions. Naturally,* $\mathrm{Cov}\{R_q^* \,|\, \theta, n\}$ *will converge to* q *for all* θ *values as* $n \to \infty$*, but very large* n *values would be necessary for a good approximation, especially for extreme values of* θ*.*

6 Further Examples

The canonical binomial and normal examples have systematically been used above to illustrate the ideas presented. In this final section a wider range of examples is presented.

6.1 Exponential data

Consider a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from an exponential distribution $\mathrm{Ex}(x \,|\, \theta) = \theta e^{-x\theta}$, and let $t = \sum_{j=1}^n x_j$.

The exponential intrinsic discrepancy loss is

$$\begin{aligned} \delta_{\boldsymbol{x}}(\theta_0, \theta) &= n \, \min[\kappa\{\theta \,|\, \theta_0\}, \, \kappa\{\theta_0 \,|\, \theta\}] \\ \kappa\{\theta_i \,|\, \theta_j\} &= \int_0^\infty \mathrm{Ex}(x \,|\, \theta_j) \log \frac{\mathrm{Ex}(x \,|\, \theta_j)}{\mathrm{Ex}(x \,|\, \theta_i)} \, \mathrm{d}x \\ &= g(\theta_i / \theta_j), \end{aligned}$$

where $g(x)$ is the *linlog* function, the positive distance

$$g(x) = (x - 1) - \log x, \tag{6.1}$$

between $\log x$ and its tangent at $x = 1$. Hence,

$$\delta_{\boldsymbol{x}}(\theta_0, \theta) = n \; \delta_x(\theta_0, \theta), \quad \delta_x(\theta_0, \theta) = \begin{cases} g(\theta_0/\theta) & \theta_0 \le \theta, \\ g(\theta/\theta_0) & \theta_0 > \theta. \end{cases}$$

A related loss function, $\ell\{\sigma_0^2, \sigma^2\} = g(\sigma_0^2/\sigma^2)$, often referred to as the *entropy loss*, was used by Brown (1968) (who attributed this to C. Stein) as an alternative to the quadratic loss in point estimation of scale parameters.

The reference prior (which here is also Jeffreys prior) is $\pi(\theta) = \theta^{-1}$, and the corresponding reference posterior is $\pi(\theta \,|\, \boldsymbol{x}) = \mathrm{Ga}(\theta \,|\, n, t) \propto \theta^{n-1} e^{-nt}$. Hence, the posterior loss $d(\theta_0 \,|\, \boldsymbol{x})$ from using θ_0 as a proxy for θ is

$$d(\theta_0 \,|\, \boldsymbol{x}) = d(\theta_0 \,|\, t, n) = n \int_0^\infty \delta_x(\theta_0, \theta) \, \mathrm{Ga}(\theta \,|\, n, t) \, \mathrm{d}\theta.$$

Figure 10 describes the behaviour of $d(\theta_o \,|\, \boldsymbol{x})$ given $n = 12$ observations, simulated from $\mathrm{Ex}(x \,|\, 2)$, which yielded $t = 4.88$. The intrinsic estimate is 2.364 (marked with a solid dot), and the intrinsic 0.95-credible interval consists of the values (1.328, 4.198) whose posterior expected loss is smaller than 1.954.

A reference parameter for this problem is $\phi(\theta) = \int \pi(\theta) \, \mathrm{d}\theta = \log \theta$, and its posterior mean may be analytically obtained as

$$\mu_\phi = \int_0^\infty \log \theta \; \mathrm{Ga}(\theta \,|\, n, t) \, \mathrm{d}\theta = \psi(n) - \log t \tag{6.2}$$

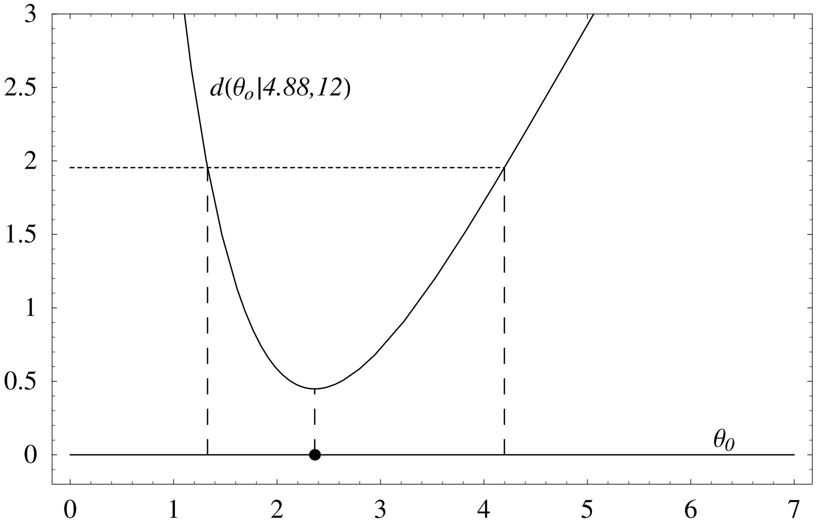

Figure 10: *Expected intrinsic loss from using* θ_0 *as a proxy for the parameter* θ *of an exponential distribution, given* $n = 12$ *observations, with* $\overline{x} = t/n = 0.407$. *The intrinsic estimate is* $\theta^* = 2.364$, *the intrinsic* 0.95*-credible region is* $(1.328, 4.198)$.

where $\psi(\cdot)$ is the digamma function. Using Equation (6.2) in Corollary 4.1, with $t = 4.88$ and $n = 12$, yields $\mu_\phi = 0.858$, and hence, $\theta^* \approx \exp[\mu_\phi] = 2.357$ very close to its exact value 2.364, even though the sample size, $n = 12$, is rather small. Moreover,

$$R^*_{0.95} \approx (\exp[\mu_\phi - 1.96/\sqrt{n}],\, \exp[\mu_\phi + 1.96/\sqrt{n}]) = (1.339,\ 4.151)$$

quite close again to the exact intrinsic region (1.328, 4.198).

In this problem, all reference posterior credible regions are *exact* frequentist confidence intervals. Indeed, changing variables, the reference posterior distribution of $\tau = t\,\theta$ is $\mathrm{Ga}(\tau\,|\,n, 1)$; on the other hand, the sampling distribution of t is $\mathrm{Ga}(t\,|\,n, \theta)$ and, therefore, the sampling distribution of $s = t\,\theta$ is $\mathrm{Ga}(s\,|\,n, 1)$. Thus, the reference *posterior* distribution of $t\,\theta$, as a function of θ, is precisely the same as the *sampling* distribution of $t\,\theta$, as a function of t; consequently, for any region $R(t) \subset \Theta$, $\Pr[\theta \in R(t)\,|\,t, n] = \Pr[\theta \in R(t)\,|\,\theta, n]$.

6.2 Uniform data

Consider a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a uniform distribution $\mathrm{Un}(x\,|\,\theta) = \theta^{-1}$, $0 \le x \le \theta$, $\theta > 0$, and let $t = \max\{x_1, \ldots, x_n\}$. Notice that this is not a regular problem, since the sample space, $\mathcal{X} = [0,\,\theta]$, depends on the parameter θ. The function $\hat{\theta} = t$ is a sufficient, consistent estimator of θ, whose sampling distribution is inverted Pareto,

$$p(t\,|\,\theta) = \mathrm{Ip}(t\,|\,n, \theta^{-1}) = n\,t^{n-1}\,\theta^{-n}, \quad 0 < t < \theta. \tag{6.3}$$

Using (6.3) and (4.5), the reference prior is immediately found to be

$$\pi(\theta) \propto p(t\,|\,\theta)\big|_{t=\theta} = \theta^{-1};$$

notice that in this non-regular problem Jeffreys rule is not applicable. The corresponding reference posterior is Pareto

$$\pi(\theta\,|\,\boldsymbol{x}) = \mathrm{Pa}(\theta\,|\,n, t) = n\,t^n\,\theta^{-(n+1)}, \quad \theta \geq t.$$

The intrinsic discrepancy loss for this model is

$$\begin{aligned} \delta_{\boldsymbol{x}}(\theta_0, \theta) &= n\,\min[\kappa\{\theta\,|\,\theta_0\},\,\kappa\{\theta_0\,|\,\theta\}] \\ \kappa\{\theta_i\,|\,\theta_j\} &= \begin{cases} \int_0^{\theta_j} \theta_j^{-1} \log[\theta_i/\theta_j]\,\mathrm{d}x = \log[\theta_i/\theta_j], & \theta_j \leq \theta_i \\ \infty & \theta_j > \theta_i \end{cases} \end{aligned}$$

Hence, $\delta_{\boldsymbol{x}}(\theta_0, \theta) = n\,|\log(\theta/\theta_0)|$, and the posterior loss $d(\theta_0\,|\,\boldsymbol{x})$ from using θ_0 as a proxy for θ is

$$d(\theta_0\,|\,\boldsymbol{x}) = d(\theta_0\,|\,t, n) = n\int_t^\infty |\log(\theta/\theta_0)|\,\mathrm{Pa}(\theta\,|\,n, t)\,\mathrm{d}\theta.$$

Figure 11 describes the behaviour of $d(\theta_0\,|\,t, n)$ given $n = 12$ observations, simulated from $\mathrm{Un}(x\,|\,2)$, which yielded $t = 1.806$. The intrinsic estimate is 1.913 (marked with a solid dot) and the 0.95 intrinsic credible interval consists of the values (1.806, 2.318) whose posterior expected loss is smaller than 2.096.

A reference parameter for this problem is $\phi(\theta) = \int \pi(\theta)\,\mathrm{d}\theta = \log\theta$, and its posterior mean may be analytically obtained as

$$\mu_\phi = \int_0^\infty \log\theta\,\mathrm{Pa}(\theta\,|\,n, t)\,\mathrm{d}\theta = (1/n) + \log t \tag{6.4}$$

For $t = 1.806$ and $n = 12$ yields $\mu_\phi = 0.674$ and $\theta^* \approx e^{\mu_\phi} = 1.963$ not too far from the exact value 1.913. Notice, however, that Theorem 4.1 cannot

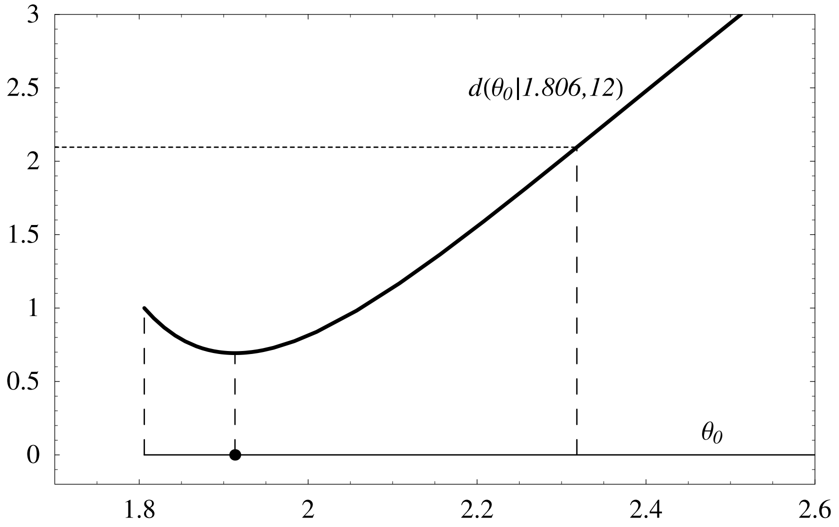

Figure 11: *Expected intrinsic loss from using* θ_0 *as a proxy for the parameter* θ *of a uniform distribution, given* $n = 12$ *observations, with* $t = \max x_j = 1.806$*. The intrinsic estimate is* $\theta^* = 1.913$*, the intrinsic* 0.95*-credible region is* $(1.806,\ 2.318)$.

be applied to this problem, since neither the sampling distribution of the consistent estimator t, nor the posterior distribution of ϕ are asymptotically normal. In fact, the posterior variance of ϕ is (exactly) $1/n^2$; this is $O(n^{-2})$ rather than $O(n^{-1})$, as obtained in regular models.

Once again, all reference posterior credible regions in this problem are *exact* frequentist confidence intervals. Indeed, changing variables, the reference posterior distribution of $\tau = \theta/t$ is Pareto, $\mathrm{Pa}(\tau \,|\, n, 1)$; on the other hand, the sampling distribution of t is inverted Pareto $\mathrm{Ip}(t \,|\, n, \theta^{-1})$ and, therefore, the sampling distribution of $s = \theta/t$ is also $\mathrm{Pa}(s \,|\, n, 1)$. Hence, the reference posterior distribution of θ/t as a function of θ is precisely the same as the sampling distribution of θ/t as a function of t and thus, for any region $R(t) \subset \Theta$, $\Pr[\theta \in R(t) \,|\, t, n] = \Pr[\theta \in R(t) \,|\, \theta, n]$.

6.3 Normal mean and variance

Consider a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a normal distribution $\mathrm{N}(x \,|\, \mu, \sigma)$, and let $\boldsymbol{\theta} = (\mu, \sigma)$ be the (bivariate) quantity of interest. The intrinsic discrepancy loss for this model is

$$\delta_{\boldsymbol{x}}\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\} \;=\; \delta_{\boldsymbol{x}}\{(\mu_0, \sigma_0), (\mu, \sigma)\}$$

$$= \quad n \min[\kappa\{(\mu_0, \sigma_0) \,|\, (\mu, \sigma)\},\ \kappa\{(\mu, \sigma) \,|\, (\mu_0, \sigma_0)\}]$$

with

$$\kappa\{(\mu_i, \sigma_i) \,|\, (\mu_j, \sigma_j\} = \frac{1}{2}\left[g\left(\frac{\sigma_j^2}{\sigma_i^2}\right) + \frac{(\mu_i - \mu_j)^2}{\sigma_i^2}\right], \tag{6.5}$$

where $g(x) = (x-1) - \log x$ is, again, the linlog function; this yields

$$\delta_{\boldsymbol{x}}\{(\mu_0, \sigma_0), (\mu, \sigma)\} = \begin{cases} n\,\kappa\{(\mu, \sigma) \,|\, (\mu_0, \sigma_0)\}, & \sigma \geq \sigma_0 \\ n\,\kappa\{(\mu_0, \sigma_0) \,|\, (\mu, \sigma)\}, & \sigma < \sigma_0. \end{cases} \tag{6.6}$$

The normal is a location-scale model and, thus (Bernardo, 2005b), the reference prior is $\pi(\mu, \sigma) = \sigma^{-1}$. The corresponding (joint) reference posterior distribution, $\pi(\mu, \sigma \,|\, \boldsymbol{x})$, is given in Equation (4.9).

The reference posterior expected intrinsic loss from using (μ_0, σ_0) as a proxy for (μ, σ) is then

$$d(\mu_0, \sigma_0 \,|\, \boldsymbol{x}) = \int_0^\infty \int_{-\infty}^\infty \delta_{\boldsymbol{x}}\{(\mu_0, \sigma_0), (\mu, \sigma)\}\, \pi(\mu, \sigma \,|\, \boldsymbol{x})\, \mathrm{d}\mu\, \mathrm{d}\sigma.$$

This is a convex surface with a unique minimum at the intrinsic estimate

$$\{\mu^*(\boldsymbol{x}), \sigma^*(\boldsymbol{x})\} = \arg \min_{\mu_0 \in \mathbb{R}, \sigma_0 > 0} d(\mu_0, \sigma_0 \,|\, \boldsymbol{x}) = \{\overline{x},\ \sigma^*(s, n)\} \tag{6.7}$$

where σ^* is of the form $\sigma^*(s, n) = k_n\, s$ and, hence, it is an affine equivariant estimator. With $n = 2$, $\sigma^*(x_1, x_2) \approx (\sqrt{5}/2)\,|x_1 - x_2|$; for moderate or large sample sizes,

$$\sigma^*(s, n) = k_n\, s \approx \sqrt{\frac{n}{n-2}}\; s. \tag{6.8}$$

Since intrinsic estimation is invariant under reparametrization, the intrinsic estimator of the variance is simply $(\sigma^*)^2 \approx n\, s^2/(n-2)$, slightly larger than both the mle estimator s^2, and the unbiased estimator $n\, s^2/(n-1)$.

Intrinsic credible regions are obtained by projecting into the (μ_0, σ_0) plane the intersections of the surface $d(\mu_0, \sigma_0 \,|\, \boldsymbol{x})$ with horizontal planes.

Figure 12 describes the behaviour of $d(\mu_0, \sigma_0 \,|\, \boldsymbol{x})$ given $n = 25$ observations, simulated from $\mathrm{N}(x \,|\, 0, 1)$, which yielded $\overline{x} = 0.024$ and $s = 1.077$. The resulting surface has a unique minimum at $(\mu^*, \sigma^*) = (0.024, 1.133)$, which is the intrinsic estimate; notice that

$$\mu^* = \overline{x}, \quad \sigma^* \approx s\sqrt{n/(n-2)} = 1.123.$$

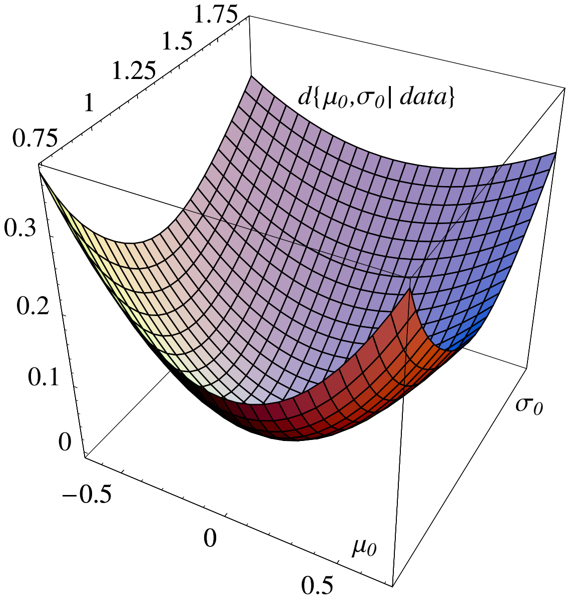

Figure 12: *Expected intrinsic loss from using* (μ_0, σ_0) *as a proxy for the parameters* (μ, σ) *of a normal distribution, given* $n = 25$ *observations, with* $\overline{x} = 0.024$ *and* $s = 1.077$.

Figure 13 represents the corresponding intrinsic estimate and contours of intrinsic q-credible regions, for $q = 0.50$, 0.95 and 0.99. For instance, $R^*_{0.95}$ (middle contour in the figure) is the set of $\{\mu_0, \sigma_0\}$ points whose intrinsic expected loss is not larger that 0.134.

Notice finally that all reference posterior credible regions in this problem are, once more, *exact* frequentist confidence intervals. Indeed, for all n, the joint reference posterior distribution of

$$\frac{\mu - \overline{x}}{\sigma/\sqrt{n}} \times \frac{n\, s^2}{\sigma^2} \tag{6.9}$$

as a function of (μ, σ) is precisely the same as its sampling distribution as a function of $(\overline{x}, s^2)$. Thus, for any region $R = R(m, s, n) \subset \mathbb{R} \times \mathbb{R}^+$, one must have $\Pr[(\mu, \sigma) \in R \,|\, m, s, n] = \Pr[(\mu, \sigma) \in R \,|\, \mu, \sigma, n]$.

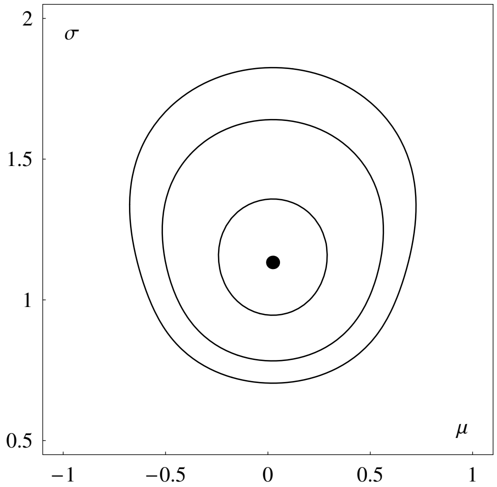

Figure 13: *Intrinsic estimate (solid dot) and intrinsic q-credible regions ($q = 0.50$, 0.95 and 0.99) for the parameters (μ, σ) of a normal distribution, given $n = 25$ observations, with $\overline{x} = 0.024$ and $s = 1.077$.*

DISCUSSION

George Casella
Department of Statistics
University of Florida, USA

1 Introduction

The vagaries of email originally sent Professor Bernardo's paper into Limbo rather than Florida, and because of that, my time to prepare this discussion was limited. As a result, I decided to concentrate on one aspect of the paper, that having to do with discrete intervals.

First, let me say that Professor Bernardo has, once again, brought us a fundamentally new way of approaching a problem, an approach that is not only extremely insightful, but also is likely to lead to even more developments in objective Bayesian analysis. The coupling of interval construction with lowest posterior loss is a very intriguing concept, and the argument for using an intrinsic loss is compelling.

There is one point about loss functions that I really like. Professor Bernardo notes that a loss $\ell\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\}$ should be measuring the distance between the models $p(\boldsymbol{x}|\boldsymbol{\theta}_0)$ and $p(\boldsymbol{x}|\boldsymbol{\theta})$, not the distance between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$, which is often irrelevant to an inference. This is an excellent principle to focus on in any decision problem. Results that are not invariant to $1-1$ transformations can sometimes be interesting in theory, but they tend to be less useful in practice.

2 Convincing Confidence Intervals

Professor Bernardo states that in the binomial problem "Conventional frequentist theory fails to provide a *convincing* confidence interval" (my italics), and then comments on the limitations imposed by the discreteness of the problem. It is not clear what a "convincing" confidence interval is - it seems to me that any confidence interval that maintains its guaranteed confidence level is convincing. It is also unclear to me, and has been for a long time, why the fact that the problem is discrete automatically brings about criticism.

The discreteness of the data is a limitation. When we impose a continuous criterion, satisfying such a criterion will often require more than the data can provide. This is not a fault of any procedure, simply a limitation of the data. The fact that in discrete problems a confidence interval cannot attain *exact* level q is not a cause for criticism.

However, what is a cause for criticism is the reliance on Bayesian intervals being approximate frequentist intervals. Although it is true that in some cases the frequentist coverage may be of the order $q + O(n^{-1})$, that O may be so big as to not be useful.

3 Binomial Confidence Intervals

Now I would like to focus on Example 5.1 (also note the companion Examples 1.1, 3.2, and 4.1). Professor Bernardo is not happy with the frequentist answer here (or anywhere, I dare say!) however, I claim that in this case the best frequentist region provides a very acceptable Bayesian region, while the objective Bayesian region fails as a frequentist region.

First of all, what is the "best" frequentist answer? To me, it is the procedure of Blyth and Still (1983) (see also Casella, 1986). This procedure works within the limitations of the discrete binomial problem to produce intervals that are not only short, but enjoy a number of other desirable properties, including equivariance. Figure 1 shows the Blyth-Still intervals along with the Bernardo intervals for the case $n = 10$.

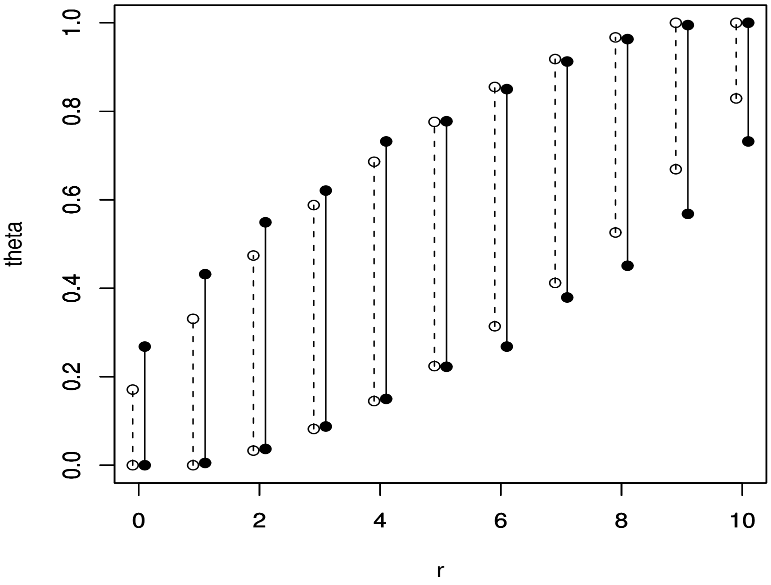

Figure 1: For $n = 10$, binomial intervals of Bernardo (dashed) and Blyth-Still (solid). The Bernardo intervals are 95% Bayesian credible intervals, and the Blyth-Still intervals are 95% frequentist intervals.

It is interesting that the intervals are so close, but we notice that the Blyth-Still intervals are a bit longer than the Bernardo intervals. Indeed, if we compare the procedures using the sum of the lengths as a measure of size, we find that the sum of the lengths of the Blyth-Still intervals is 5.20, while that of the Bernardo intervals is 4.53. However, one of the criteria that the Blyth-Still intervals satisfy is that, among level q confidence intervals, they minimize the sum of the lengths. Therefore, we know that the Bernardo intervals cannot maintain level q and, indeed, from Bernardo's Figure 9 we see that this is indeed the case. Even though the Bernardo intervals are approximate frequentist intervals, the approximation is really quite poor. The nominal 95% interval can have coverage probability as low as 83% (reading off the graph) which is quite unacceptable. Moreover, the fluctuations in the coverage probability are quite large, ranging from a low of 83% to a high of 100%

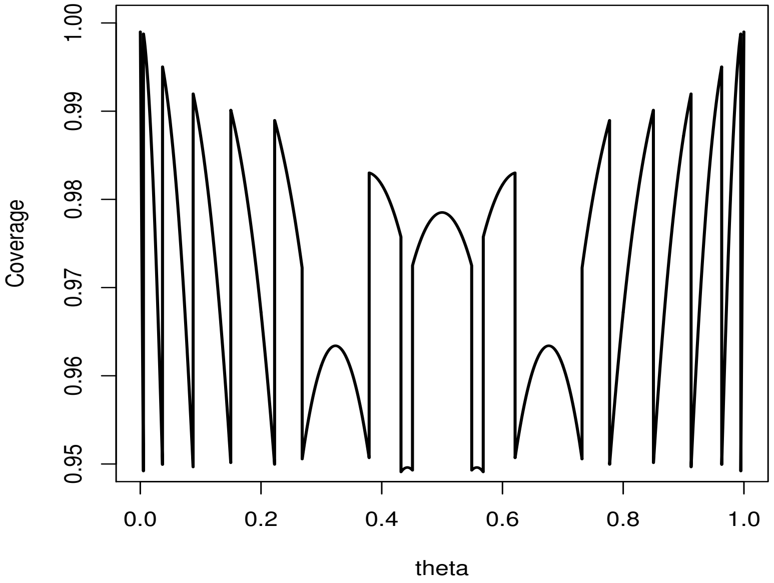

Figure 2: *For* $n = 10$, *coverage probabilities of the* 95% *Blyth-Still intervals.*

The frequentist intervals, although not convincing to Professor Bernardo, do a fine job of controlling the coverage probability within the constraints of a discrete problem. As an illustration, compare Bernardo's Figure 9 with Figure 2, showing the coverage probability of the Blyth-Still 95% interval. Although there is fluctuation in the probabilities, they are above 95%, making a true confidence interval, and the range of probabilities ranges only from 95% to 100%, displaying much less variability than the Bernardo intervals.

Finally, lets look at how the Blyth-Still intervals fare as Bayesian credible regions. Using the reference prior, we can produce Table 1. There we see that they are, indeed, 95% credible intervals. Although the credible probabilities are not exactly 95% for each value of r, they are uniformly greater than r, varying in the range $.951 - .989$.

Table 1: *Credible probabilities of the* 95% *Blyth-Still confidence intervals*

r	0	1	2	3	4	5
Prob	.989	.979	.971	.959	.964	.951

r	6	7	8	9	10	
Prob	.964	.959	.971	.979	.989	

What to conclude from all of this? As a long-time frequentist, it is supremely gratifying to see the wide development of objective Bayesian inference, which is defined by Professor Bernardo as a statistical analysis that only depends on the model and the observed data. With one more small step, we might include the sample space and then objective Bayesians will attain the ultimate objective goal of being frequentists!

But, on a more serious note, we see that in "objective" inference, there is a desire to have a procedure perform well on Bayesian (post-data) and frequentist (pre-data) criteria. What my illustration was supposed to demonstrate is that one can construct an acceptable objective Bayesian procedure not just by starting from a Bayesian derivation and then checking frequentist properties, but also by starting from a frequentist derivation and then checking Bayesian properties.

Edward I. George
Department of Statistics
University of Pennsylvania, USA

Let me begin by congratulating Professor Bernardo on a deep methodological contribution that is based on important recent foundational considerations. I must admit that I was skeptical when I began to read this paper, thinking that this would simply be a new recipe for credible regions with little in the way of compelling justification. Much to my surprise, the proposed approach has both a compelling motivation and turns out, at least in some cases, to dovetail nicely with frequentist procedures that are routinely used. Going further, I would recommend these intrinsic credible regions as the new default in the absence of reliable prior information.

As I understand it, the general goal of the objective Bayes approach is to devise default Bayesian procedures that require minimal, if any, subjective input. Towards this end, the basic problem in this paper is to come up with an objective Bayesian methodology for the construction of credible regions for an unknown parameter function value $\boldsymbol{\theta}(\boldsymbol{\omega})$.

At the heart of the objective Bayesian approach is the choice of prior, and in this regard, a key development has been the default reference prior approach put forward by Professor Bernardo and his collaborators. These attractive reference priors are an essential ingredient for the proposed construction of intrinsic credible regions in so far as they are used to obtain the reference posterior. However, the central issue tackled in this paper is how to extract from the reference posterior a unique set of values which are in a meaningful sense "closest" to $\boldsymbol{\theta}(\boldsymbol{\omega})$. As Professor Bernardo correctly points out, HPD regions are unappealing because of a lack of reparametrization invariance, and posterior quantile regions may easily exclude sets of highest posterior concentration.

Instead, Professor Bernardo proposes using a Lowest Posterior Loss (LPL) region, effectively a neighborhood of the minimum posterior expected loss estimate. I much prefer such an approach because it treats the region estimation problem as an extension of the decision theoretic treatment of the point estimation problem. Although a companion coverage report of $(1-\alpha)\%$ probability, frequentist or Bayesian, is valuable, I think it has incorrectly been given too much primacy as a construction criterion. I believe a practitioner is best served by a region containing point estimates that are superior to point estimates outside the region, which is precisely what LPL is about. I don't think it is wise to sacrifice this property simply to gain a more accurate coverage estimate.

Having settled on LPL, the problem becomes one of choosing a loss function that is essentially objective. For this purpose, Professor Bernardo argues convincingly that attention must be restricted to the so-called intrinsic losses, losses that are parametrization invariant and so depend only on discrepancies between models. I agree. Further, divergence measures are natural candidates for such losses. Especially attractive is the KL divergence which as Professor Bernardo notes is invariant under sample space transformation, see also George and McCulloch (1993). However, the lack of symmetry of the KL divergence is problematic for the construction of LPL regions. By proposing instead to use the intrinsic discrepancy loss, Professor Bernardo gets to "have his cake and eat it too". The intrinsic discrepancy essentially symmetrizes the KL divergence without relinquishing many of its wonderful features. And I much prefer this symmetrization to the standard alternative approach of a weighted sum of the two directed KL divergences.

The examples in Section 3 nicely illustrate how intrinsic discrepancies between distributions can depend dramatically upon different regions in the parameter space. Indeed, Example 3.1 shows how clearly the intrinsic discrepancy reveals regions of asymptotic equivalence and nonequivalence between the binomial and Poisson models. To me, this highlights the need to avoid arbitrary loss functions for the construction of LPL regions. Further investigation into other potential uses of the intrinsic discrepancy seems to be a fertile new research direction.

Thus, Professor Bernardo arrives at his definition of an intrinsic credible region – namely an LPL region based on intrinsic discrepancy loss and the appropriate reference prior – very reasonable and well motivated. But I then was astonished to see what came next. Applied to several important examples, namely the normal mean, the exponential, the uniform, and the normal mean and variance, the intrinsic credible regions are all also exact frequentist confidence intervals. These intrinsic credible regions they not only contain a best set of point estimates, but their coverage can be conveyed in an objectively meaningful way. I suspect that further investigation of this agreement may shed valuable new light on the always fascinating interface between Bayesian and frequentist methods. I wonder if Professor Bernardo can pinpoint the essential reason behind this matching property of intrinsic credible regions in these cases, and if he has any sense of how broadly it will it occur?

The frequentists coverage properties of the intrinsic credible regions for the discrete binomial distribution are not as nice, but this is fundamentally a general problem of all interval estimation reports for discrete distributions, see Brown et al. (2001). My sense is that a different type of report is needed in such cases, for example, see Geyer and Meeden (2005). In any case, the asymptotic agreement with frequentist coverage is reassuring.

It is clear that when reliable prior subjective information is unavailable, a Bayesian analysis must turn to non-subjective considerations. For this purpose Professor Bernardo has made wonderful use of the invariance properties of information theoretic measures. It is interesting that such invariance is typically lacking in fully subjective Bayesian methods.

In closing, I must apologize to Professor Bernardo for being so positive about his paper, as it hardly gives him anything to argue about in his rejoinder. But that is the price for having such a good idea.

Javier Girón
Department of Statistics
University of Málaga, Spain
Elías Moreno
Department of Statistics
University of Granada, Spain

The advantage of a unifying approach to the basic inference problems, as the one carried out by Professor Bernardo in this and previous papers on intrinsic estimation and testing, (see Bernardo, 2001) is always welcomed and, in this respect, we cannot help but congratulate Professor Bernardo for his proposal.

Highest posterior density regions (HPD's) are, as Professor Bernardo correctly asserts in his paper, a tool we employ to summarize a given posterior density $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$. Sometimes the mode, mean and standard deviation are also reported.

A natural question to be answered before giving credit to the *intrinsic credible regions* proposed in the paper is the following: do we need to replace HPD's with another region? Professor Bernardo asserts that HPD's are not *coherent* in the sense of not being invariant under reparametrization. Therefore, the consequence for him is that a different region is needed. But, not so fast!

Given a one-to-one differentiable transformation $\boldsymbol{\alpha} = g(\boldsymbol{\theta})$ the posterior density of $\boldsymbol{\alpha}$ is

$$\pi(\boldsymbol{\alpha} \,|\, \boldsymbol{x}) = \pi(g^{-1}(\boldsymbol{\alpha}) \,|\, \boldsymbol{x})|\mathbf{J}|, \tag{1}$$

where $|\mathbf{J}|$ stands for the determinant of the Jacobian of the inverse mapping. Then, HPD's for the parameter $\boldsymbol{\alpha}$ are obtained from the posterior distribution $\pi(\boldsymbol{\alpha} \,|\, \boldsymbol{x})$. What is wrong with this?

Professor Bernardo claims that the HPD's are not *coherent* because the q-HPD for $\boldsymbol{\alpha}$ does not coincide with the g-transformation of the q-HPD for $\boldsymbol{\theta}$. This is obviously due to the fact that the Jacobian in Equation (1) is not constant unless g be linear. We do not think this *coherent* notion to be of such a fundamental nature, and we find such a requirement rather artificial.

In fact, q-HPD regions are a good summary of the posterior density and they are defined in the same spirit as the likelihood sets notion. These sets have been proved to enjoy very nice properties (Cano et al., 1987; Piccinato, 1984; Wasserman, 1989).

But intrinsic credible regions rely more on the properties of the intrinsic loss discrepancy than on the form of the posterior. Thus, the computation of intrinsic credible regions appears as a somewhat contrived artifact to assure coherence, i.e. invariance, rather than a means to show off nice characteristics of the posterior density for a given parametrization. Further, the metric or scale of the expected intrinsic loss is in units of information δ^*, while intrinsic credible regions are measured in a probabilistic scale q; consequently, in order to compute intrinsic credible regions the expected intrinsic loss has to be *calibrated* in terms of the probabilistic content q.

Though for one-dimensional parameters when the expected intrinsic loss is pseudoconvex and the posterior density is unimodal HPD's and intrinsic credible regions are intervals, contours in more than one dimension, obtained from the expected intrinsic loss may be very different from the ones obtained from the posterior density, whatever the parametrization, specially in higher dimensions. This might be very disturbing, and it is the (sometimes very high) price one has to pay to preserve invariance; for this reason, we believe that credible regions should capture the properties of the posterior density not those of the expected intrinsic loss. We note that the above comments apply to any other loss function we would consider appropriate for the estimation problem, as the differences between the contours of the posterior risk and the posterior density can be very substantial.

One point of some concern is the fact that, as shown in the binomial Example 4.1 and the uniform example in Section 6.1, the reference posterior expected intrinsic loss displayed in Figures 7 and 11, respectively, though they are convex they are not increasing. This means that for very small values of q the intrinsic credible intervals do not contain values in a small neighborhood of 0, thus ruling out a set of points with highest posterior density; furthermore, this problem also holds for any monotonic transformation of the parameter $\boldsymbol{\theta}$.

It might be argued that intrinsic credible regions do only make sense for the usual large values of q, say 0.95 and 0.99, but from a formal standpoint the behavior of the intrinsic credible regions should be consistent whatever the value of q.

On the other hand, as the two statistical problems of testing a sharp hypothesis of the form $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and computing a credible region in the parameter space $\boldsymbol{\Omega}$ both use the reference posterior expected intrinsic loss in a similar way, it is apparent that there exists a duality between the two problems in the same sense as in the classical approach and in the Bayesian one advocated by Lindley (1970, pp. 56–69). In fact, to reject the null $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is equivalent to see if the reference posterior expected intrinsic loss $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})$ is greater that some suitable positive value δ^*, where δ^* is the expected utility of accepting the null hypothesis when it is true. This is obviously equivalent to rejecting the null whenever $\boldsymbol{\theta}_0$ does not belong to the credible region R^*_q for some q which depends on δ^* defined by

$$R^*_q = \{\boldsymbol{\theta}; d(\boldsymbol{\theta} \,|\, \boldsymbol{x}) < \delta^*\}.$$

In some sense the computation of δ^*, which is carried out conditioning on the null hypothesis, resembles that of computing a p-value in frequentist testing.

Thus, from Bernardo's approach, intrinsic testing of sharp null hypothesis and intrinsic credible regions are equivalent procedures as there is a one to one mapping between δ^*'s and q's. Further, this mapping depends entirely on the reference posterior expected intrinsic loss, thus differing from Lindley's approach to significance testing.

While this approach has many advantages —the most important one being that no new prior but the reference prior is to be used for estimation and testing and, in some sense, provides a Bayesian simple mechanism for significance testing from a statistical viewpoint—, the issue of practical significance seems to be missing in this approach as we believe that point or interval estimation is quite a different statistical problem than that of testing sharp nulls, and this in turn means that using a prior different from the reference one which may take into account the sharp null to be tested makes sense in this setting.

Daniel Peña
Departamento de Estadística
Universidad Carlos III de Madrid, Spain

This article presents an objective way to obtain confidence intervals for a parameter. In Bayesian inference this problem has not received much attention and the Bayesian literature usually has stressed that the posterior probability for the parameter incorporates all the information about it and from this posterior probability it is straightforward to obtain a q-credible interval, that is, a region on the support of the random variable with has a probability equal to q. Among all the infinite regions that can be built a natural criterion is that values inside the interval must have higher probability than values outside, but in order to apply this rule it is well known, see for instance Lindley (1970, p. 35), that we have to decide which parameter is to be used, as this rule is not invariant under one-to-one transformations of the parameter.

In this article an objective solution is proposed to solve this ambiguity. The procedure is to use a reference prior and a reference loss function and one key contribution in this paper is the use of the intrinsic discrepancy loss, introduced by Bernardo and Rueda (2002) for hypothesis testing. The author has to be congratulated for proposing a clear and objective way to solve this ambiguity in building credible regions.

The intrinsic discrepancy is defined as the minimum of $\kappa(p_2 \,|\, p_1)$ and $\kappa(p_1 \,|\, p_2)$ where

$$\kappa(p_2 \,|\, p_1) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx \geq 0$$

is the Kullback-Leibler information in favor of p_1 against p_2. This measure is invariant under reparametrization and also with respect to one to one transformations of the variable. It is also additive for independent observations. These properties are shared with the divergence $\kappa(p_2 \,|\, p_1) + \kappa(p_1 \,|\, p_2)$, introduced by Kullblack and Leibler. However it has the advantage over the later that it is well defined when one of the two measures of information is infinite. To avoid this problem, Kullback (1968) assumed that the two probability measures were absolutely continuous with respect to each other, so that the two integrals could exist. The intrinsic discrepancy does not need this constrain and therefore it has a wider applicability. However, in the application considered in this paper, building credible regions, it is to be expected that the two probability measures used verify Kullback conditions, that is they are absolutely continuous with respect to each other, and therefore this property may not be crucial. Assuming that in most cases both discrepancies can be used, which are the advantages for building credible regions of the intrinsic discrepancy with respect to the standard Kullblack Leibler divergence?, how different the results would be of using one instead of the other?

The derivation of intrinsic credible regions may be complicated and may require numerical integration and thus an interesting contribution from the point of view of applications is Theorem 4.1, where simple asymptotic approximations are obtained. This result will facilitate the use of the presented theory in practice.

My final comments on this interesting and though-provoking paper are on three directions. First, it would be useful to have a better understanding of the advantages of the proposed approach over the standard HPD regions. Suppose that I have a HPD region for θ. Would it be possible to compute a measure of the maximum loss that we may have if we translate this HPD for θ to build a credible interval for a one-to-one transformation of the parameter ϕ? From my point of view in order to convince people to use these ideas is important to provide some bounds of the advantages that we may obtain with respect to conventional methods.

Second, how this ideas can be extended for building prediction credible regions for future values of the observed random variable of interest? As in the Bayesian approach parameters are random variables, I suppose the extension is straightforward but it would be interesting to have some comments on this topic.

Third, how these ideas can be extended for dependent observations as, for instance, time series or spatial data? When the data have some dependency structure prediction of the observed random variable is usually the key problem, and an approach to develop prediction intervals which does not depend on the parametrization could be very appealing. For time series a popular measure of information is the mutual information or relative entropy (Joe, 1989) given by

$$K(x,y) = \int\int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy,$$

which is nonnegative and it is zero only if the two variables are independent. This measure has been used to build a test of independence (Robinson, 1991), identifying lags in the relationship in nonlinear time series modeling (Granger and Lin, 1994) or building measures of general dependency among vector random variables (Peña and van der Linde, 2005), among others applications. Taking $p_1(x) = f(x, y)$ and $p_2(x) = f(x)f(y)$ the intrinsic discrepancy between dependency and independency of the two random variables x and y can be defined and this idea might be used for credible regions and hypothesis testing in time series. It would be interesting to explore this possibility.

Finally I would like to congratulate the author for a very interesting piece of research.

Judith Rousseau and Christian P. Robert
CEREMADE, Université Paris Dauphine
and CREST, INSEE, France

In this paper, Professor Bernardo presents a unified and structured objective approach to (decisional) statistical inference, based on information theoretic ideas he has used previously to define reference priors. He focusses here on the estimation of credible regions, keeping those values of the parameter that are the least costly rather than the most probable, as in HPD regions. This is an interesting and novel approach to an efficient construction of credible regions when lacking a decision-theoretic basis. As noted in Casella et al. (1993, 1994) (see also Robert, 2001, Section 5.5.3, for a review), the classical decision-theoretic approaches to credible regions are quite behind their counterpart for point estimation and testing and incorporating loss perspectives in credible sets was then suggested in Robert and Casella (1994).

1 On invariance: link with HPD regions

A possible drawback of HPD regions, in particular in objective contexts, is their lack of invariance under reparameterization as was pointed out by Professor Bernardo. Obviously, HPD regions are defined in terms of a volume-under-fixed-coverage loss and they do minimize the volume among

q-credible regions. The lack of invariance hence stems from the lack of invariance in the definition of the volume, which is based on the Lebesgue measure for the considered parametrization θ. Therefore, simply considering a different type of volume based on an invariant measure would result in an HPD region that is invariant under reparameterization. A natural invariant measure in this setup is Jeffreys' measure, due to its geometric and information interpretations (among others). The resulting HPD region is thus constructed as the region C that minimizes

$$\int_C \sqrt{i(\theta)} d\theta, \quad \text{u.c.} \quad P^\pi[C|X] \geq q. \tag{1}$$

This region also corresponds to the transform of the (usual) HPD region constructed using the reference parametrization as defined by Professor Bernardo.

Note that, in the above construction, there is absolutely no need in having the prior be Jeffreys prior and this construction could be used in (partially) informative setups. It is also interesting to note that, in regular cases, the above HPD region is asymptotically equivalent to the intrinsic credible region of Professor Bernardo. Which of both approaches is the most appealing is probably a question of taste or depends on how they will be used.

On a more philosophical basis, we think that invariance is less compelling an argument for (credible) regions than for point estimations. Indeed, while it is difficult to sell to a customer that the estimator of $h(\theta)$ is not necessarily the transform $h(\hat{\theta})$ of the estimator $\hat{\theta}$ of θ, the transform of a crebible region does remain a credible region, even though it is not always the optimal region. Moreover, invariance under reparameterization should be weighted against shape poor modifications. Indeed, if we impose that the credible region C_h on $h(\theta)$ is the transform by h of the credible region C_{id} on θ, we get exposed to strange shapes for less regular functions h! For instance, if the transform h is not monotonic (but still one-to-one), it is possible to obtain the transform of a credible interval as a collection of several disjoint intervals, always a puzzling feature! Connexity (and maybe to some extent convexity) should be part of the constraints on a credible region.

2 Asymptotic coverage : matching properties

Under regularity properties, the HPD region defined by (1) is a second order matching region for any smooth prior π, in the sense that its frequentist coverage is equal to its posterior coverage to the order $O(n^{-1})$. Third order coverage does not necessarily apply for Jeffreys' prior, though (see Datta and Mukerjee, 2004 or Rousseau, 1997). As Bernardo's intrinsic credible region is asymptotically equivalent to the HPD region defined by (1) there is a chance that second order matching is satisfied, which would explain the good small sample properties mentioned in the paper. In particular, the perturbation due to using the intrinsic loss, compared to using the posterior density, is of order $O(n^{-1})$, so second order asymptotics should be the same between (1) and the intrinsic credible region.

Investing further the higher order matching properties of this credible region would be worthwhile though. Regarding the discrete case, however, things are more complicated than what was mentioned by Professor Bernardo since there is usually no matching to orders higher than $O(n^{-1/2})$ or sometimes $o(n^{-1/2})$ for higher dimensional cases. Whether reference posterior q-credible regions provide the best available solution for this particular problem is somehow doubtful as there are many criteria which could reasonably be considered for comparing credible regions or their approximations in the discrete case, see Brown et al. (2002).

3 Computations

Adopting this approach to credible set construction obviously makes life harder than computing HPD regions: while HPD regions do simply require the derivation of a posterior level ϱ for the set $\{\theta : \pi(\theta|x) \geq \varrho\}$ to have coverage q, an intrinsic credible set involves the intrinsic loss—not easily computed outside exponential families—, the posterior intrinsic loss—possibly integrated over a large dimensional space—, the posterior coverage of the corresponding region and at last the bound on $d(\theta|x)$ that garantees q coverage. In large dimensional settings or outside exponential frameworks, the tasks simply seems too formidable to be contemplated, especially given that standard numerical features like convexification cannot be taken for granted since the credible region is not necessarily convex or even connected.

Rejoinder by J. M. Bernardo

I am most grateful to the editors for inviting so many fine statisticians to comment on this paper, and to the discussants for their kind remarks and thought-provoking comments. I will try to provide a personalized answer to each of them.

1 Casella

I am obviously flattered by the opening paragraphs of Professor Casella's comment, and I am glad to read that he appreciates the importance of requiring that statistical procedures should be invariant under reparametrization.

Given his mainly frequentist standpoint, it is not surprising that he finds convincing a confidence interval even if this cannot exactly obtain a required level. He claims that this is a consequence of the limitation imposed by the discreteness of the data and, mathematically, this is certainly true. My point, however, is that, since the parameter is continuous, one should expect to be able to provide region estimates for the parameter in a continuous manner, and this obviously requires a Bayesian approach. It may be argued that what scientists need is a set of parameter values which, *given the available data*, may be expected to be close to the true value; the average properties under sampling of the procedure are certainly worth investigating but, I believe, they should not be the overwhelming consideration.

We should all be grateful to Professor Casella for the detailed comparison between the solution to interval estimation of a binomial parameter proposed in the paper and that of Blyth-Still (that Professor Casella considers the best frequentist answer), for this provides a nice example where the practical implications of foundational issues may usefully be discussed. I should first mention an important foundational difference: while the Bayesian approach provides general procedures, which may be applied without modification to any particular example, good frequentist procedures often require a great deal of ingenuity to taylor the procedure to the specific needs of the example ("adhockeries" in de Finetti, 1970, terminology); this is clearly demonstrated by the long history of frequentist confidence intervals in the binomial case.

The practical differences between the two solutions compared mirror their very different foundational basis. Indeed, if one takes the frequentist view that the criterion is to have *at least* 95% coverage with a minimum length, then Blyth-Still solution does a very good job and, as one could expect, this produces longer intervals than the Bayesian solution, with posterior probability larger than 0.95. If, on the other hand, one takes the Bayesian view that the criterion is to have precisely 0.95 posterior probability, one has shorter intervals with an *average* 95% coverage. Professor Casella finds a particular 83% coverage unacceptable from a frequentist viewpoint if 95% was the desired level; I similarly find unacceptable from a Bayesian viewpoint a posterior probability 0.989 if 0.95 was required. I suspect that the two methods would give very close *numerical* answers if either the Bayesian procedure were set to a credible level equal to the *average* confidence level reached by the frequentist procedure or, alternatively, if the frequentist procedure were set to a confidence level equal to the *minimum* coverage attained by the Bayesian procedure.

The main difference however, is not numerical but foundational; it does not lie with the numerical differences (in many other examples, as illustrated in the paper, the numbers are precisely equal) but with their interpretation. The main point, I believe, is whether a scientist is best served by a given interval estimate and the knowledge, that had he obtained the data he had *not*, the resulting intervals would have contained the true value 95% of the time, or by a (possibly different) interval and the knowledge the he is entitled to a 0.95 measure of uncertainty, in a $(0, 1)$ scale, that this *particular* interval contains the true value. I firmly believe that most scientists, if given the choice, would prefer the second scenario.

That said, I must applaud Professor Casella's suggestion that frequentist statisticians should check the Bayesian properties of their proposals. As this paper illustrates, good Bayesian and frequentist solutions are often numerically very close, and both paradigms (and their fertile interface) should make part of any decent education in mathematical statistics.

2 George

I must warmly thank Professor George for his excellent review of the motivation and contents of the paper. As he states in the last paragraph of his comment, it hardly leaves me anything to argue about!

Professor George wonders about the essential reasons behind the matching properties of intrinsic credible regions. The set of examples I have chosen to include possibly has an over-representation of exact matching cases. Indeed, except for the binomial case (where, as in any problem with discrete data, exact matching is impossible), all the other examples show exact matching in the sense that the frequentist coverage of q-credible regions is exactly q for any sample size. In all these examples, this is due to the existence of a pivotal quantity whose sampling distribution as a function of the data is precisely the same as its reference posterior distribution as a function of the parameters. I suspect the existence of such a pivotal quantity is the basic condition for exact matching to exist; related pioneering papers are Lindley (1958) and Barnard (1980). Whether or not the reference posterior of the pivotal quantity (whenever this exists) is *always* the same as its sampling distribution is an interesting open problem. I would think that this is probably true (under appropriate regularity conditions) in one-dimensional problems, but I doubt that such a result would generalize to multivariate settings. More work is certainly needed in that direction.

A superficial reading of the exact matching properties may however lead to think that, when pivotal quantities do exist, intrinsic credible regions give the same numerical result than conventional frequentist confidence sets, but this is certainly *not* the case. Indeed, q-credible regions may well have exact q-coverage and yet be numerically very different from the common frequentist q-confidence sets. For instance, the conventional q-confidence interval for the normal variance when the mean is unknown, based on a probability centred interval on the χ^2_{n-1} distribution of the pivotal quantity ns^2/σ^2, is

$$C_q = \left[\frac{ns^2}{Q_{n-1}\{(1+q)/2)\}}\,,\,\frac{ns^2}{Q_{n-1}\{(1-q)/2\}}\right] \tag{1}$$

where $Q_\nu\{p\}$ is the p-quantile of a χ^2_{n-1} distribution. On the other hand, extending the results in (Bernardo, 2005a) to region estimation, the q-intrinsic credible region is the set R_q^* of the σ_i^2 values such that

$$\int_{R_q^*} \pi(\sigma^2 \,|\, n, s^2)\, \mathrm{d}\sigma^2 = q, \quad \forall \sigma_i^2 \in R_q^*,\ \forall \sigma_j^2 \notin R_q^*,\ d(\sigma_i^2 \,|\, n, s^2) \le d(\sigma_j^2 \,|\, n, s^2)$$

where $\pi(\sigma^2 \,|\, n, s^2)$, the reference posterior of σ^2, is an inverted gamma $\mathrm{Ig}(\sigma^2 \,|\, (n-1)/2,\, ns^2/2)$,

$$d(\sigma_i^2 \,|\, n, s^2) = \frac{1}{2} \int_0^\infty \delta\left(\frac{\sigma_i^2\, \tau}{ns^2}\right) \chi^2(\tau \,|\, n-1)\, \mathrm{d}\tau, \tag{2}$$

and $\delta(\theta) = g(\theta)$ if $\theta < 1$, $\delta(\theta) = g(\theta^{-1})$ if $\theta > 1$, with $g(x) = (x-1) - \log x$. Using the results in Theorem 4.1, this may approximated by

$$R_q^* \approx \exp\left[\left\{\left(\log[\frac{ns^2}{2}] - \psi(\frac{n-1}{2})\right) \pm z_q \sqrt{\psi'(\frac{n-1}{2})}\right\}\right], \tag{3}$$

where $\psi(\cdot)$ is the digamma function, and z_q is the standard normal quantile of order $(1+q)/2$. Using Stirling to approximate the polygamma functions this further reduces to

$$R_q^* \approx \exp\left[\left\{\left(\frac{1}{n-1} + \log\frac{n\,s^2}{n-1}\right) \pm z_q\, \frac{\sqrt{2\,n}}{n-1}\right\}\right]. \tag{4}$$

With a simulated sample of size $n = 20$ from a $\mathrm{N}(x \,|\, 0, 1)$ distribution, which yielded $\overline{x} = 0.069$ and $s^2 = 0.889$, the conventional 0.95-confidence set for σ^2 is $C_{0.95} = (0.5376, 1.9828)$, while the exact intrinsic interval is $R_{0.95}^* = (0.5109, 1.8739)$, the approximation (3) yields $(0.5104, 1.8840)$, and (4) gives $(0.5102, 1.8812)$. As one would expect, the differences between confidence sets and credible regions increase as the sample size decreases. To show an extreme case, with only two observations, $x_1 = -1$ and $x_2 = 1$, the 0.95-confidence set is $C_{0.95} = (0.398, 2037)$, while the intrinsic interval is the far better centred $R_{0.95}^* = (0.004, 509)$; even in this extreme case the approximation (3) is useful, yielding $(0.092, 554)$, while (4) gives $(0.108, 274)$. Again, the dual behaviour of the pivotal quantity ns^2/σ^2 guarantees in this example that the coverage probability of a Bayesian credible interval R_q, and reference posterior probability of a frequentist confidence interval C_q are both *exactly* q, whatever the sample size.

3 Girón and Moreno

Professors Girón and Moreno question the use of the emotionally charged adjective "incoherent" to refer to the lack of invariance under one-to-one reparametrization shown by HPD intervals, which was the terminology I used in the first draft of this paper. They are probably right, and I have replaced this by simply "non-invariant" is the final version. That said, I still believe invariance under (irrelevant) one-to-one reparametrization should be a requirement for *any* statistical procedure, an in particular, this should be a property of appropriate region estimators.

Professors Girón and Moreno suggest that the use of an invariant loss function may be a contrived artifact to achieve invariance rather than a means to show the properties of the posterior. I disagree. They seem to miss the point that *any* feature of the posterior which is not invariant under reparametrization is completely illusory, since the parametrization is arbitrary. Points with relatively large posterior density in one parametrization may well correspond to points with relatively low posterior density in another. The need for invariant interval regions stems from the fact that the main objective of a region estimate, which is to give the scientist a set of values of the parameter of interest which (given the available data) could reasonably be expected to be close to the true value, *requires* invariance. Would anyone be happy to report a set of, say, credible speed values for a remote galaxy given available measurements to a group of scientists working in speed units, and a *different* set to another group working in a log scale?

I have chosen to define the low posterior loss regions in terms of credible regions to facilitate comparison with conventional interval estimates. However, there is no need to specify the threshold level δ^* indirectly, in probability terms, as implied by a posterior probability content. Indeed, δ^* may directly be specified in terms of the maximum expected loss one is prepared to suffer. If the intrinsic discrepancy loss is used, this simply means to exclude from the region estimate those parameter values which label models leading to expected log-likelihood ratios against them larger than δ^*. As Professors Girón and Moreno correctly point out this is closely related to precise hypothesis testing and, in hypothesis testing, it leads to a different (and I would claim better) solution than the conventional Bayes factor approach. There is no space here however to discuss this important issue, and the interested reader is referred to Bernardo and Rueda (2002) and Bernardo (2005b).

Professors Girón and Moreno find disturbing the fact that the contours obtained from LDL regions may be very different, specially in higher dimensions, from those of HPD regions. It is difficult to argue without any specific example in mind, but what I would definitely find disturbing is to include in a region estimate parameter values with large expected losses. The crucial difference between focusing attention on probability density rather than expected loss is well illustrated by one of their final comments. The show concern on the fact, illustrated by Example 4.1, that high density parameter values, such as zero in a binomial situation when no successes have been observed, may be excluded from an intrinsic credible interval with an small credible level. Thus, in the $n = 10$, $r = 0$ case discussed in the paper, the 0.50-intrinsic credible interval is $(0.006, 0.070)$, and this excludes the parameter values close to 0, although the posterior density of θ is monotonically decreasing from 0. Notice, however, that very few statisticians would consider the point estimator $\tilde{\theta} = 0$, which is both the maximum likelihood estimate and the posterior mode in this case, anything but useless; thus, most would use some sort of correction to obtain an strictly positive point estimate of θ and, indeed, the intrinsic point estimate in this case (see Example 4.1) is $\theta^* = 0.031$. It is then natural to expect that if an small set of *good* estimates of θ is desired, these should concentrate in a small neighbourhood of the optimal choice, which is θ^*. The important point to notice is that values around θ^* have a *lower expected loss* than the more likely values around zero. In practice, this means that, after observing $r = 0$ successes in n trials, if one were to act *as if* the true model were $\mathrm{Bi}(r \,|\, \tilde{\theta}, n)$ it would be safer to act as if $\tilde{\theta}$ were close to θ^*, the intrinsic estimate, than to act as if $\tilde{\theta}$ were close to 0, the posterior mode. This is, I believe, an eminently sensible conclusion.

4 Peña

Professor Peña wonders whether the results would be very different if, in those cases where both directed divergences are finite, the conventional symmetric logarithmic divergence (already suggested by Jeffreys)

$$\ell_J(p_1, p_2) = \kappa\{p_1 \,|\, p_2\} + \kappa\{p_2 \,|\, p_1\}$$

were used instead of the intrinsic loss in Definition 3.1; I would expect the two solutions in that case to be pretty similar. However, as illustrated by the uniform data example of Section 6.2, there are many interesting problems where one of the directed divergences is infinite, and I believe one should strive for theories which are as generally applicable as possible. As illustrated by the Poisson approximation to the Binomial (see Example 3.1), the intrinsic discrepancy may be applied to important statistical problems where Jeffreys divergence cannot.

As stated, Definition 4.2 may be applied to any type of data, including those with *dependent* observations. Indeed, $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ stands for the *joint* distribution of the observed data $\boldsymbol{x}$ (although in all the examples considered $\boldsymbol{x}$ has been assumed to be a random sample from some underlying distribution). If Professor Peña, a known expert in dependent data, decided to try out a particular example with dependent observations, we would all be grateful.

Professor Peña makes the interesting suggestion of providing a measure of the expected loss from using a HPD region instead of the optimal LDL region. The posterior expected loss from using any particular value $\boldsymbol{\theta}_i$ as a proxy for the unknown value of the parameter is given by $d(\boldsymbol{\theta}_i \,|\, \boldsymbol{x})$, in Equation (4.1). Thus, an upper bound of the expected loss increase from using a region estimate $R_q(\boldsymbol{x}, \boldsymbol{\Theta})$ rather than the optimal region $R^*(\boldsymbol{x}, \boldsymbol{\Theta})$ would be the (necessarily positive) quantity

$$\Delta(R_q) = \sup_{\boldsymbol{\theta}_i \in R_q} d(\boldsymbol{\theta}_i \,|\, \boldsymbol{x}) - \sup_{\boldsymbol{\theta}_i \in R_q^*} d(\boldsymbol{\theta}_i \,|\, \boldsymbol{x}).$$

Simple approximations to $\Delta(R_q)$ should be available from the results of Theorem 4.1. A possible use of $\Delta(R_q)$ would be to check the possible losses associated to the use of HPD regions in alternative parametrizations. It follows from Theorem 4.1 that under regularity conditions, the best choice would be the reference parametrization since, in that case, intrinsic credible regions are approximate (and often exact) HPD regions.

Derivation of credible sets for prediction is certainly a natural extension of the proposed theory. This is not directly contemplated in Definition 4.2 but, as Professor Peña suggests, the main ideas may be indeed applied. In a prediction situation, the loss from using a predictive density $p_x(\cdot)$ as a function of the value x eventually observed may be argued to be of the form

$$\ell\{p_x(\cdot), x\} = -\,a \log p_x(x) + b, \quad a > 0, \quad b \in \mathbb{R},$$

for, under regularity conditions, the logarithmic scoring rule is the only proper local scoring rule (Bernardo, 1979a). The best possible predictive density is obviously the actual model $p(x \,|\, \boldsymbol{\theta})$. Hence, as a function of $\boldsymbol{\theta}$, the loss suffered from predicting a particular value x_0 would be of the form

$$\ell\{x_0, \boldsymbol{\theta}\} = -\,a \log p(x_0 \,|\, \boldsymbol{\theta}) + b, \quad a > 0, \quad b \in \mathbb{R}.$$

To set an origin for the loss scale, let $x_{\boldsymbol{\theta}}$ be some good estimate of x given $\boldsymbol{\theta}$, say the mean, the mode or the median of $p(x \mid \boldsymbol{\theta})$, which is arbitrarily given zero loss. In this case, $\ell\{x_{\boldsymbol{\theta}}, \boldsymbol{\theta}\} = -\,a \log p(x_{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) + b = 0$, and solving for b yields $b = a \log p(x_{\boldsymbol{\theta}} \mid \boldsymbol{\theta})$. Hence, as a function of $\boldsymbol{\theta}$, the loss suffered from predicting any other value x_0 would be

$$\ell\{x_0, \boldsymbol{\theta}\} = a \log \frac{p(x_{\boldsymbol{\theta}} \mid \boldsymbol{\theta})}{p(x_0 \mid \boldsymbol{\theta})}\,, \quad a > 0. \tag{5}$$

The corresponding reference posterior expected loss,

$$l(x_0 \mid \boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \ell\{x_0, \boldsymbol{\theta}\}\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}, \tag{6}$$

is invariant under both one-to-one reparametrization and one-to-one transformation of the observable x. I would suggest that a function of the form (5) is an appropriate loss function for prediction. Using the expected loss (6) in Definition 2.1, the corresponding lower posterior loss (LPL) q-credible predictive region would be a subset $R_q(\boldsymbol{x}, \mathcal{X})$ of $\mathcal{X}$ such that

$$\int_{R_q} p_x(x \mid \boldsymbol{x})\, \mathrm{d}x = q, \qquad \forall x_i \in R_q, \quad \forall x_j \notin R_q, \quad l(x_i \mid \boldsymbol{x}) \le l(x_j \mid \boldsymbol{x}).$$

For example, in the exponential example of Section 6.1, with the conditional mean $x_\theta = \mathrm{E}[x \mid \theta] = \theta^{-1}$ used to set the origin of the loss scale, the reference posterior expected loss (6) is easily found to be $d(x_0 \mid t, n) = x_0\, n/t - 1$, where t is the sum of the observations. Since this is an increasing function of x_0, the LPL q-credible predictive interval would be of the form $R_q = (0, a)$, where $a = a(q, t, n)$ is such that $\int_0^a p(x \mid t, n)\, \mathrm{d}x = q$, and $p(x \mid t, n)$ is the reference posterior predictive density,

$$p(x \mid t, n) = \frac{\Gamma(n+1)}{\Gamma(n)} \frac{t^n}{(t+x)^n}\,, \quad x > 0.$$

Analytical integration and some algebra yields the q-predictive interval

$$R_q = \bigl(0,\ [(1-q)^{-1/n} - 1]\, t\bigr).$$

For the numerical illustration considered in Section 6.1, where $n = 12$ and $t = 4.88$, the LPL 0.95-credible predictive interval (which in this case is also HPD) is $R_{0.95}(\boldsymbol{x}, \mathcal{X}) = (0,\ 1.384)$.

In his final comment, Professor Peña refers the the relative entropy, that is the directed logarithmic divergence of the product of the marginals from the joint distribution, as a sensible measure of dependence. This is a particular case (Bernardo, 2005b) of the *intrinsic dependency*,

$$\min\left[\int_{\mathcal{X}}\int_{\mathcal{Y}} p(\boldsymbol{x},\boldsymbol{y})\log\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x},\int_{\mathcal{X}}\int_{\mathcal{Y}} p(\boldsymbol{x})p(\boldsymbol{y})\log\frac{p(\boldsymbol{x})p(\boldsymbol{y})}{p(\boldsymbol{x},\boldsymbol{y})}\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}\right]$$

which reduces to the relative entropy (the first integral above) under regularity conditions, but may be seen to behave better in non-regular cases. Professor Peña suggests the use of a dependency measure as the basis for a prediction loss function. This is an interesting idea, well worth exploring.

5 Rousseau and Robert

Professors Rousseau and Robert describe a solution to invariant region estimation that, when applicable, is simple and elegant. As they point out, this is the usual HPD region obtained using the reference parametrization and thus, their proposal is asymptotically equivalent to ours. Notice, however, that there are many important non-regular problems where Jeffreys prior does not exist and hence, their method could not be used.

They correctly point out that there is no need in their construction to use Jeffreys' as a formal prior and that their construction could be used in (partially) informative setups. We note in passing that this is also the case with intrinsic credible regions: as stated in Definition 2.1, q-credible lowest posterior loss regions may be found conditional to a posterior probability q obtained from any desired prior. Indeed, their construction may be formally seen a particular case or lowest posterior loss regions where the (invariant) loss function is taken to be the volume based on Jeffreys' measure.

Professors Rousseau and Robert argue that in region estimation invariance should be weighted against shape poor modifications, and suggest that connexity should be part of the constraints on a credible region. I disagree. For instance, a non-connected region might be the only sensible option if, say, the posterior distribution is strongly bimodal. The particular parametrization of the problem is irrelevant and thus, as a basic foundational point, the procedure should be independent of the parametrization. It is certainly true that connected, convex regions may be easier to understand but, precisely because the procedure is invariant under reparametrization, one is free to choose that parametrization where the required regions look better. This will often be the reference parametrization where, as mentioned above, intrinsic regions will be nearly HPD.

Professors Rousseau and Robert refer to the matching properties of both their suggestion and the (asymptotically equivalent) intrinsic regions. While it is certainly nice to know that, asymptotically, the expected proportion of q-credible regions containing the true value is q, I believe that too much emphasis on numerical coincidence with confidence regions is misplaced. Indeed, there is a large class of problems (Gleser and Hwang, 1987), which includes for instance the region estimation of the ratio of normal means, where frequentist confidence regions may be both useless and misleading, and therefore, one does *not* want to approximate these.

I certainly share Professors Rousseau and Robert concern with computational issues: this is why I invested some effort in deriving approximate solutions. However, I believe that one should derive what the optimal procedure should be, and then try to find clever ways to obtain either numerical solutions or analytical approximations to the optimal procedure, rather than using a simple alternative (say a quadratic loss) just because it it easier to compute, even if it may be shown to be less than appropriate for the problem considered.

References

Barnard, G. A. (1980). Pivotal inference and the Bayesian controversy (with discussion). In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds., *Bayesian Statistics 1*, pp. 295–318. University Press, Valencia.

Berger, J. O. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *Journal of the American Statististical Association*, 95:1269–1276.

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statististical Association*, 84:200–207.

Berger, J. O. and Bernardo, J. M. (1992a). On the development of reference priors (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 4*, pp. 35–60. Oxford University Press, Oxford.

Berger, J. O. and Bernardo, J. M. (1992b). Ordered group reference priors with applications to a multinomial problem. *Biometrika*, 79:25–37.

Berger, J. O. and Bernardo, J. M. (1992c). Reference priors in a variance components problem. In P. K. Goel and N. S. Iyengar, eds., *Bayesian Analysis in Statistics and Econometrics*, pp. 323–340. Springer–Verlag, Berlin.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statististical Association*, 91:109–122.

Bernardo, J. M. (1979a). Expected information as expected utility. *The Annals of Statistics*, 7:686–690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society. Series B*, 41:113–147. Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.), Edward Elgar, Brookfield, VT, 1995, pp. 229–263.

Bernardo, J. M. (1997). Non-informative priors do not exist (with discussion). *Journal of Statistical Planning and Inference*, 65:159–189.

Bernardo, J. M. (2001). Un programa de síntesis para la enseñanza universitaria de la estadística matemática contemporánea. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales (España)*, 95(1–2):87–105.

Bernardo, J. M. (2005a). Intrinsic point estimation of the normal variance. In S. K. Upadhyay, U. Singh, and D. K. Dey, eds., *Bayesian Statistics and its Applications*. Anamaya Pub, New Delhi. In press.

Bernardo, J. M. (2005b). Reference analysis. In D. Dey and C. R. Rao, eds., *Bayesian Thinking, Modeling and Computation*, vol. 25 of *Handbook of Statistics*. North Holland, Amsterdam. In press.

Bernardo, J. M. and Juárez, M. (2003). Intrinsic estimation. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., *Bayesian Statistics 7*, pp. 465–476. Oxford University Press, Oxford.

Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: Inference on the ratio of multinomial parameters. *Journal of the Royal Statistical Society. Series D. The Statistician*, 47:1–35.

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*. John Wiley & Sons, Chichester.

Blyth, C. R. (1986). Approximate binomial confidence limits. *Journal of the American Statististical Association*, 81:843–855.

Blyth, C. R. and Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association*, 78:108–116.

Brown, L. (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *Annals of Mathematics and Statistics*, 39:29–48.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statistical Science*, 16:101–133.

Brown, L. D., Cai, T. T., and DasGupta, A. (2002). Confidence intervals for a binomial proportion and Edgeworth expansion. *The Annals of Statistics*, 30(1):160–201.

Burdick, R. K. and Graybill, F. A. (1992). *Confidence Intervals on Variance Components*. Marcel Dekker, New York.

Cano, J. A., Hernández, A., and Moreno, E. (1987). A note on maximized likelihood sets. *European Journal of Operational Research*, 32:291–293.

Casella, G. (1986). Refining binomial confidence intervals. *The Canadian Journal of Statistics*, 14:113–129.

Casella, G., Hwang, J. T., and Robert, C. P. (1993). A paradox in decision-theoretic set estimation. *Statistica Sinica*, 3:141–155.

Casella, G., Hwang, J. T., and Robert, C. P. (1994). Loss function for set estimation. In J. O. Berger and S. S. Gupta, eds., *Statistical Decision Theory and Related Topics, V*, pp. 237–252. Springer Verlag, New York.

Datta, G. S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*, vol. 178 of *Lecture Notes in Statistics*. Springer Verlag, New York.

Datta, G. S. and Sweeting, T. (2005). Probability matching priors. In D. Dey and C. R. Rao, eds., *Bayesian Thinking, Modeling and Computation*, vol. 25 of *Handbook of Statistics*. North Holland, Amsterdam. In press.

de Finetti, B. (1970). *Teoria delle Probabilità*. Einaudi, Turin. English translation as *Theory of Probability* in 1974, John Wiley & Sons, Chichester.

Eberly, L. E. and Casella, G. (2003). Estimating Bayesian credible intervals. *Journal of Statistical Planning and Inference*, 112:115–132.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *Journal of the American Statististical Association*, 82:171–200.

George, E. I. and McCulloch, R. E. (1993). On obtaining invariant prior distributions. *Journal of Statistical Planning and Inference*, 37:169–179.

Geyer, C. J. and Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and p-values (with discussion). *Statistical Science*. To appear.

Gleser, L. and Hwang, J. T. (1987). The nonexistence of $100(1-\alpha)$ confidence sets of finite expected diameter in errors-in-variable and related models. *The Annals of Statistics*, 15:1351–1362.

Granger, C. and Lin, J.-L. (1994). Using the mutual information coefficient to identify lags in nonlinear models. *The Journal of Time series Analysis*, 15(4):371–384.

Gutiérrez-Peña, E. (1992). Expected logarithmic divergence for exponential families. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and

A. F. M. Smith, eds., *Bayesian Statistics 4*, pp. 669–674. Oxford University Press, Oxford.

Guttman, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian*. Griffin, London.

Hahn, G. J. and Meeker, W. Q. (1991). *Statistical Intervals*. John Wiley & Sons, Chichester.

Hartigan, J. A. (1966). Note on the confidence prior of Welch and Peers. *Journal of the Royal Statistical Society. Series B*, 28:55–56.

Hartigan, J. A. (1983). *Bayes Theory*. Springer–Verlag, Berlin.

Joe, H. (1989). Relative entropy measures of multivariate dependence. *The Journal of American Statistical Association*, 405:157–164.

Juárez, M. (2004). *Métodos Bayesianos Objetivos de Estimación y Contraste de Hipótesis*. PhD Thesis, Universidad de Valencia, Spain.

Kullback, S. (1968). *Information Theory and Statistics*. Dover, New York, 2nd ed. Reprinted in 1978, Peter Smith, Gloucester, MA.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematics and Statistics*, 22:79–86.

Lindley, D. V. (1958). Fiducial distribution and Bayes' Theorem. *Journal of the Royal Statistical Society. Series B*, 20:102–107.

Lindley, D. V. (1970). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference*. Cambridge University Press, Cambridge.

Peña, D. and van der Linde, A. (2005). General measures of variability and dependence for multivariate continuous distribution. Manuscript submitted for publication.

Piccinato, L. (1984). A Bayesian property of the likelihood sets. *Statistica*, 3:367–372.

Robert, C. P. (1996). Intrinsic loss functions. *Theory and Decision*, 40:192–214.

Robert, C. P. (2001). *The Bayesian Choice*. Springer Verlag, New York, 2nd ed.

Robert, C. P. and Casella, G. (1994). Distance penalized losses for testing and confidence set evaluation. *Test*, 3(1):163–182.

Robinson, P. (1991). Consistent nonparametric entropy-based testing. *Review of Economic Studies*, 58:437–453.

Rousseau, J. (1997). Expansions of penalized likelihood ratio statistics and consequences on matching priors for HPD regions. Tech. rep., CREST, INSEE, Paris.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons, New York. Second edition in 1972, Dover, New York.

Schervish, M. J. (1995). *Theory of Statistics*. Springer–Verlag, Berlin.

Wasserman, L. (1989). A robust Bayesian interpretation of likelihood regions. *The Annals of Statistics*, 17:1387–1393.

Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on intervals of weighted likelihoods. *Journal of the Royal Statistical Society. Series B*, 25:318–329.