Distribuciones Iniciales Objetivas: Presente y Futuro

José M. Bernardo
Universidad de Valencia, Spain
jose.m.bernardo@uv.es

Workshop sobre Métodos Bayesianos
Universidad Complutense, Madrid
7 de Noviembre de 2008

The Concept of Probability

- A Bayesian approach is firmly based on *axiomatic foundations*.
  Mathematical need to describe by probabilities *all* uncertainties:
  Parameters *must* have a (*prior*) probability distribution,
  assumed to describe available information about their values.
  This probability distribution is *not* a description
  of their variability (they are *fixed unknown* quantities),
  but a description of the *uncertainty* about their true values.
  Consideration of replications not required, and often irrelevant.
- Tentatively accept a *formal* model. which describes the probabilistic relationship between data and quantities of interest.
  Model is suggested by informal *descriptive* evaluation.
  Conclusions always *conditional* on model assumptions.

Objective Bayesian Statistics

- Very important particular case:
  No relevant objective initial information.
- Includes scientific and industrial reporting,
  and public decision making.
- Prior distribution based *only* on explicit model assumptions:
  This is *Objective Bayesian Statistics*.
- Main research effort to theoretically derive the *objective* prior
  from the assumed statistical model.
- Accepted procedures use mathematical information theory:
  Reference prior, reference analysis.

Notation

- The functions $p(\boldsymbol{x})$, $p(\boldsymbol{\theta})$ are *probability* densities (or mass) functions of *observables* $\boldsymbol{x} \in \mathcal{X}$ or *parameters* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

  Special densities with specific notation:

  $\mathrm{N}(x \,|\, \mu, \sigma^2)$, $\mathrm{St}(x \,|\, \mu, \sigma, \alpha)$, or $\mathrm{Ga}(\theta \,|\, \alpha, \beta)$.

- *Model* generating $\boldsymbol{x} \in \mathcal{X}$, $\mathcal{M} \equiv \{\, p(\boldsymbol{x} \,|\, \boldsymbol{\theta}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$

  Data set $\boldsymbol{x} \in \mathcal{X}$. *Sample space* $\mathcal{X}$, of arbitrary structure.

  Quantity of interest $\phi = \phi(\boldsymbol{\theta}) \in \Phi \subset \Re$

  Alternatively, $\mathcal{M} \equiv \{\, p(\boldsymbol{x} \,|\, \phi, \boldsymbol{\omega}), \boldsymbol{x} \in \mathcal{X}, \phi \in \boldsymbol{\Phi}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$

  in terms of quantity of interest and nuisance parameters.

- *Posterior* $p(\phi \,|\, \boldsymbol{x}) \propto \int_{\boldsymbol{\Omega}} p(\boldsymbol{x} \,|\, \phi, \boldsymbol{\omega})\, p(\phi, \boldsymbol{\omega})\, d\boldsymbol{\omega}$ combines information from data $\boldsymbol{x}$ with prior information.

Reference Priors and Reference Posteriors

- *Reference prior* $\pi_\phi(\phi, \boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P}) = \pi_\phi(\phi, \boldsymbol{\omega})$
  This is that formal prior which, among all candidate priors $p(\phi, \boldsymbol{\omega}) \in \mathcal{P}$ may be expected to have a minimal effect, relative to data from $\mathcal{M}$ on the posterior inference about ϕ.
- *Reference posterior*
  This is obtained by stantard use of probability theory (Bayes theorem and appropriate integration) as $\pi(\phi \,|\, \boldsymbol{x}) \propto \int_{\boldsymbol{\Omega}} p(\boldsymbol{x} \,|\, \phi, \boldsymbol{\omega})\, \pi_\phi(\phi, \boldsymbol{\omega})\, d\boldsymbol{\omega}$.
- The reference posterior encapsutates all relevant information about the quantity of interest ϕ provided by data $\boldsymbol{x}$, under the assumptions implied by $\mathcal{M}$.

Divergence Measures

□ Hellinger distance

$h\{p_1, p_2\} = \int_{\mathcal{X}} \left(\sqrt{p_1(\boldsymbol{x})} - \sqrt{p_2(\boldsymbol{x})}\right)^2 d\boldsymbol{x}$

It is a metric, but it is *not additive*;

If $p_i(\boldsymbol{x}) = \prod_{j=1}^n q_i(x_j), \quad h\{p_1, p_2\} \neq n\, h\{q_1, q_2\}$

□ Logarithmic divergence

The logarithmic divergence (Kullback-Leibler) $\kappa\{p_2 \,|\, p_1\}$ of a density $p_2(\boldsymbol{x})$, $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_2$, from a true density $p_1(\boldsymbol{x})$, $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_1$, is $\kappa\{p_2 \,|\, p_1\} = \int_{\mathcal{X}_1} p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\, d\boldsymbol{x}$, (provided this exists).

- $\kappa\{p_2 \,|\, p_1\} \geq 0$ is zero iff, $p_2(\boldsymbol{x}) = p_1(\boldsymbol{x})$, a.e.
  it is *invariant* under one-to-one transformations of $\boldsymbol{x}$, and
  it is *additive*: $\kappa\{p_1 \,|\, p_2\} = n\, h\{q_1 \,|\, q_2\}$
- But $\kappa\{p_1 \,|\, p_2\}$ is *not symmetric* and *diverges* if $\boldsymbol{\mathcal{X}}_2 \subset \boldsymbol{\mathcal{X}}_1$.

☐ Intrinsic discrepancy

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

The *intrinsic discrepancy* $\delta\{p_1, p_2\}$ is *symmetric*,
non-negative, and zero iff, $p_1 = p_2$, a.e.
invariant under one-to-one transformations of $\boldsymbol{x}$,
additive: If $p_i(\boldsymbol{x}) = \prod_{j=1}^{n} q_i(x_j)$, $\quad \delta\{p_1, p_2\} = n\, \delta\{q_1, q_2\}$

• With strictly nested supports the intrinsic discrepancy is still well defined: If, strictly, $\mathcal{X}_i \subset \mathcal{X}_j$, then $\delta\{p_i, p_j\} = \kappa\{p_j \,|\, p_i\}$.

☐ Intrinsic convergence of distributions

• *Intrinsic Convergence.* A sequence of probability densities $\{p_i(\boldsymbol{x})\}_{i=1}^{\infty}$ converges *intrinsically* to $p(\boldsymbol{x})$ if (and only if) the intrinsic divergence between $p_i(x)$ and $p(x)$ converges to zero. *i.e.*, iff $\lim_{i\to\infty} \delta(p_i, p) = 0$.

Permissible Priors

□ Proper approximation of improper priors

- Objective Bayesian methods often use *improper priors*, non-negative functions $\pi(\boldsymbol{\theta})$ such that $\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ is not finite.
- If $\pi(\boldsymbol{\theta})$ is an improper prior function, $\{\boldsymbol{\Theta}_i\}_{i=1}^{\infty}$ is a sequence approximating $\boldsymbol{\Theta}$, such that $\int_{\boldsymbol{\Theta}_i} \pi(\boldsymbol{\theta}) < \infty$, and $\{\pi_i(\boldsymbol{\theta})\}_{i=1}^{\infty}$, are the proper priors obtained by *renormalizing* $\pi(\boldsymbol{\theta})$ within each of the $\boldsymbol{\Theta}_i$'s, then

  For all data $\boldsymbol{x}$ with likelihood $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, the sequence of posteriors $\{\pi_i(\boldsymbol{\theta} \mid \boldsymbol{x})\}_{i=1}^{\infty}$, with $\pi_i(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})\, \pi_i(\boldsymbol{\theta})$
  converges intrinsically to $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$.
- This justifies the formal use of improper prior functions.

Class of permissible priors

• A positive function $\pi(\boldsymbol{\theta})$ is an *permissible* prior function for model $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}),\ \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ if :

(i) for all $\boldsymbol{x} \in \mathcal{X}$, $\int_{\boldsymbol{\Theta}} p(\boldsymbol{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$,

(ii) for some sequence $\{\boldsymbol{\Theta}_i\}_{i=1}^{\infty}$ such that $\lim_{i \to \infty} \boldsymbol{\Theta}_i = \boldsymbol{\Theta}$, and $\int_{\boldsymbol{\Theta}_i} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$,

$$\lim_{i \to \infty} \int_{\mathcal{X}} p_i(\boldsymbol{x})\, \delta\{\pi_i(\boldsymbol{\theta} \mid \boldsymbol{x}),\ \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\}\, d\boldsymbol{x} = 0,$$

where $\pi_i(\boldsymbol{\theta})$ is the renormalized restriction of $\pi(\boldsymbol{\theta})$ to $\boldsymbol{\Theta}_i$, $\pi_i(\boldsymbol{\theta} \mid \boldsymbol{x})$ is the corresponding posterior, $p_i(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}_i} p(\boldsymbol{x} \mid \boldsymbol{\theta})\, \pi_i(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$. (Strong intrinsic convergence).

• All proper priors are permissible, but improper priors may or may not be permissible, even if they are arbitrarily close to proper priors.

Intrinsic Association

- The intrinsic association $\alpha\{p_{\boldsymbol{xy}}\}$ between two random vectors $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$ with joint density $p_{\boldsymbol{xy}}$ and marginals $p_{\boldsymbol{x}}$ and $p_{\boldsymbol{x}}$ is the intrinsic discrepancy $\alpha\{p_{\boldsymbol{xy}}\} = \delta\{p_{\boldsymbol{x}\,\boldsymbol{y}},\, p_{\boldsymbol{x}} p_{\boldsymbol{x}}\}$ between their joint density and the product of their marginals.

  It is a non-negative invariant measure of association between two random vectors, which vanishes if they are independent.

- The *coefficient of association*,

$$\gamma\{p_{\boldsymbol{xy}}\} = 1 - \exp[-2\alpha\{p_{\boldsymbol{xy}}\}]$$

  is a general measure of stochastic dependence on $[0, 1]$.

- In particular, if p_{xy} is bivariate normal, with coefficient of correlation ρ, then $\alpha\{p_{xy}\} = -\frac{1}{2}\log(1 - \rho^2)$, and $\gamma\{p_{xy}\} = \rho^2$.

Expected Information

- The expected intrinsic information $I\{p_{\boldsymbol{\theta}} \,|\, \mathcal{M}\}$ which one observation from model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}),\ \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ may be expected to provide about $\boldsymbol{\theta}$ when the prior is $p(\boldsymbol{\theta})$ is defined as the intrinsic dependence $\delta\{p_{\boldsymbol{x}\,\boldsymbol{\theta}},\ p_{\boldsymbol{x}}\, p_{\boldsymbol{\theta}}\}$ between $\boldsymbol{x}$ and $\boldsymbol{\theta}$, where $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})$, and $p(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$.

☐ Properties of the expected intrinsic information

- For a fixed model $\mathcal{M}$, the expected intrinsic information $I\{p_{\boldsymbol{\theta}} \,|\, \mathcal{M}\}$ is a concave, positive functional of the prior $p(\boldsymbol{\theta})$.
- Under appropriate regularity conditions $I\{p_{\boldsymbol{\theta}} \,|\, \mathcal{M}\}$ reduces to Shannon's expected information (cf. Lindley, 1956)

$$I\{p_{\boldsymbol{\theta}} \,|\, \mathcal{M}\} = I_s\{p_{\boldsymbol{\theta}} \,|\, \mathcal{M}\} = \int_{\mathcal{X}} p(\boldsymbol{x}) \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, \log \frac{p(\boldsymbol{\theta} \,|\, \boldsymbol{x})}{p(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\, d\boldsymbol{x}.$$

□ Intuitive Basis for Reference Priors

• Given model $\mathcal{M}$, the intrinsic information $I\{p_{\boldsymbol{\theta}} \,|\, \mathcal{M}\}$ measures, as a functional of the prior $p(\boldsymbol{\theta})$, the information about the value of $\boldsymbol{\theta}$ which one observation $\boldsymbol{x} \in \mathcal{X}$ may be expected to provide.

• The stronger the prior knowledge described by $p(\boldsymbol{\theta})$, the smaller the information the data may be expected to provide; conversely, weak initial knowledge about $\boldsymbol{\theta}$ (relative to the information which data from $\mathcal{M}$ could possibly provide) will correspond to large expected information from the data generated from $\mathcal{M}$.

• Define the *missing information* about the quantity of interest as that which *infinite* independent replications of the experiment could possible provide.

• Define the *reference prior* as that which *maximizes the missing information about the quantity if interest*.

Reference Distributions

• Given model $\{p(\boldsymbol{x} \mid \theta), \boldsymbol{x} \in \mathcal{X}, \theta \in \Theta \subset \Re\}$, consider $I\{p_\theta \mid \mathcal{M}^k\}$ the information about θ which may be expected from k conditionally independent replications of the original setup when the prior is $p(\theta)$. As $k \to \infty$, this would provide any *missing information* about θ, and the functional $I\{p_\theta \mid \mathcal{M}^k\}$ will approach the missing information about the value of θ associated with the prior p_θ.

• Let $\pi_k(\theta) = \pi_k(\theta \mid \mathcal{M}, \mathcal{P})$ (if it exists) be the unique proper prior which maximizes $I\{p_\theta \mid \mathcal{M}^k\}$ in the class $\mathcal{P}$ of strictly positive candidate prior distributions (compatible with accepted assumptions on the value of θ).

• If the sequence $\{\pi_k(\theta)\}$ exists, the *reference prior* $\pi(\theta) = \pi(\theta \mid \mathcal{M}, \mathcal{P})$ is defined as a limit of the sequence of priors $\{\pi_k(\theta)\}$.

Formal Definition

- In general, however, the supremum of $I\{p_\theta \mid \mathcal{M}^k\}$ is not necessarily attained at a proper prior $\pi_k(\theta)$ within the candidate class $\mathcal{P}$, and a more general definition is needed.

- **Definition**. (Berger, Bernardo and Sun, 2008). Consider model $\mathcal{M} \equiv \{p(\boldsymbol{x} \mid \phi), \boldsymbol{x} \in \mathcal{X}, \phi \in \boldsymbol{\Phi} \subset \Re\}$ and class of priors $\mathcal{P}$. The positive function $\pi(\phi) = \pi(\phi \mid \mathcal{M}, \mathcal{P})$ is a reference prior for model $\mathcal{M}$ given $\mathcal{P}$ if it is a permissible prior such that, for some sequence $\{\Phi_i\}_{i=1}^\infty$ with $\lim_i \Phi_i = \Phi$ and $\int_{\Phi_i} \pi(\phi)\, d\phi < \infty$,

$$\forall p \in \mathcal{P}, \quad \lim_{k \to \infty} \{I\{\pi_i \mid \mathcal{M}^k\} - I\{p_i \mid \mathcal{M}^k\}\} \geq 0$$

where $\pi_i(\phi)$ and $p_i(\phi)$ are the restrictions of $\pi(\phi)$ and $p(\phi)$ to Φ_i.

Explicit Expression

• **Theorem**. Consider $\mathcal{M} \equiv \{ p(\boldsymbol{x} \mid \phi), \boldsymbol{x} \in \mathcal{X}, \phi \in \boldsymbol{\Phi} \subset \Re\}$ and the class $\mathcal{P}_0$ of all regular priors for ϕ. Let $\boldsymbol{t}_k = \boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) \in \mathcal{T}$ be a sufficient statistic for $\mathcal{M}^k$, $h(\phi)$ be any positive function such that, for sufficiently large k, $c_k = \int_\Phi p(\boldsymbol{t}_k \mid \phi)\, h(\phi)\, d\phi < \infty$. Define

$$f_k(\phi) = \exp\left\{ \int_{\mathcal{T}} p(\boldsymbol{t}_k \mid \phi)\, \log \pi_k(\phi \mid \boldsymbol{t}_k)\, d\boldsymbol{t}_k \right\},$$

where $\pi_k(\phi \mid \boldsymbol{t}_k) = p(\boldsymbol{t}_k \mid \phi)\, h(\phi)/c_k$, and let

$$f(\phi) = \lim_{k\to\infty} f_k(\phi)/f_k(\phi_0), \quad \text{for any} \quad \phi_0 \in \Phi.$$

Then, if $f(\phi)$ is a permissible prior, any function of the form $\pi(\phi \mid \mathcal{M}, \mathcal{P}_0) = c\, f(\phi)$ is a reference prior for model $\mathcal{M}$.

☐ Explicit form under regularity conditions

- **Corollary 1**. Let $\tilde{\phi}_k = \tilde{\phi}(\boldsymbol{t}_k)$ be a consistent, asymptotically sufficient estimator of ϕ (often the MLE), $\hat{\phi}$.

  For large k, $\pi_k(\phi) \approx \exp[\mathrm{E}_{\tilde{\phi}_k \,|\, \phi}\{\log \pi_k(\phi \,|\, \tilde{\phi}_k)\}]$

  As $k \to \infty$, $\mathrm{E}_{\tilde{\phi}_k \,|\, \phi}\{f(\tilde{\phi}_k)\}$ converges to $f(\phi)$.

  Hence, $\pi(\phi \,|\, \mathcal{M}, \mathcal{P}_0) = \pi(\phi \,|\, \tilde{\phi}_k)|_{\tilde{\phi}_k = \phi}$

- Under regularity conditions, the posterior distribution of ϕ is asymptotically Normal, $\mathrm{N}\{\phi \,|\, \hat{\phi}, [n\, i(\hat{\phi})]^{-1/2}\}$, where

  $i(\phi) = -\mathrm{E}_{x \,|\, \phi}[\partial^2 \log p(\boldsymbol{x} \,|\, \phi)/\partial\phi^2]$

  is Fisher's information function.

  Hence, $\pi(\phi \,|\, \mathcal{M}, \mathcal{P}_0) = i(\phi)^{1/2}$ (Jeffreys' rule).

- Thus, Jeffreys rule is a particular case of a reference prior, only appropriate for one-parameter regular continuous problems.

□ Posterior summaries.

Point estimates,

Credible intervals (or regions) and

and Test of hypothesis are

all partial inferential statements,

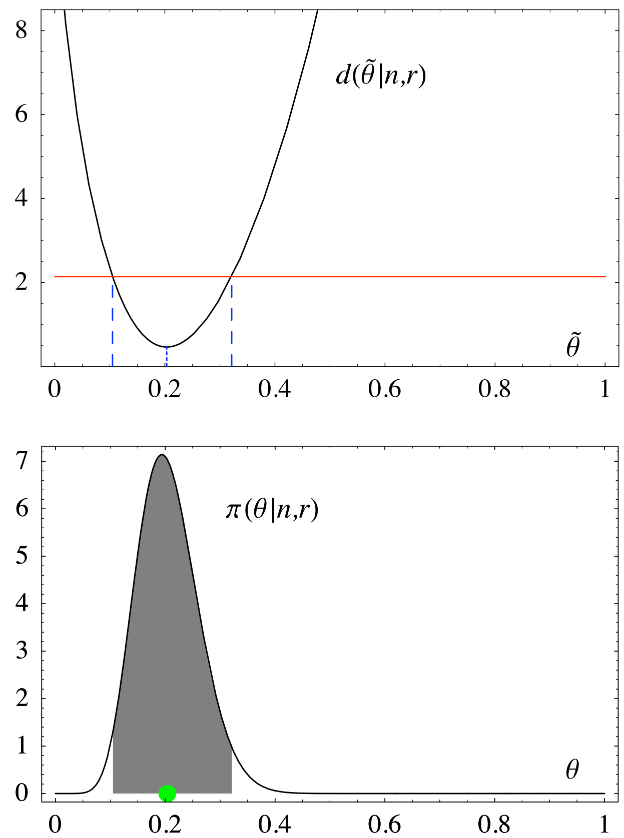
- They are easily derived

  from the reference posterior.

- They are defined using an

  appropriate, information-based,

  invariant loss function.



1. Discrete Parameter Spaces

- *Conventional Solutions: Bayes and Laplace*

  - Model $\{\, p(\boldsymbol{z} \,|\, \theta) = \prod_{i=1}^{n} p(x_i \,|\, \theta), \quad x \in \mathcal{X}, \quad \theta \in \Theta = \{\theta_1, \theta_2, \ldots\} \,\}$
    Prior $\Pr(\theta_j) \propto 1, \quad j \in \mathcal{J} \subset \mathcal{N}$
    Posterior $\Pr(\theta_j \,|\, \boldsymbol{z}) \propto p(\boldsymbol{z} \,|\, \theta_j), \quad \theta \in \Theta\}$
    Predictive $p(x \,|\, \boldsymbol{z}) = \sum_{j \in \mathcal{J}} p(x \,|\, \theta_j) \Pr(\theta_j \,|\, \boldsymbol{z})$.

- *Reference prior with finite parameter space (Maximum entropy)*

  - If $\Theta = \{\theta_1, \ldots, \theta_m\}$ is finite,
    $$\pi_\theta = \underset{p_\theta \in \mathcal{P}}{\text{Arg Max}}\; H[p_\theta], \qquad H[p_\theta] = -\sum_{j=1}^{m} p_\theta(\theta_j) \log p_\theta(\theta_j)$$
    If $\mathcal{P} = \{\text{All distributions over } \Theta\}$, $\pi_\theta(\theta_j) = 1/m, \;\; j = 1, \ldots, m$.
  - If Θ is not finite, the existence of the reference prior depends on the choice of the class $\mathcal{P}$ of permissible priors.

- *Uniform priors are often inappropriate*

  - **Example 1.** *Binomial model*

    In a Binomial $\mathrm{Bi}(r \mid n, \theta)$ model with both parameters unkown, the use of a prior of the form $p(n, \theta) = \mathrm{Be}(\theta \mid \frac{1}{2}, \frac{1}{2})$, uniform on n and Jeffreys on θ, produces an *improper* posterior for θ.

  - **Example 2.** *Hypergeometric model*

    The hypergeometric model $\mathrm{Hy}(r \mid n, R, N)$ converges, as $N \to \infty$, to a binomial model $\mathrm{Bi}(r \mid n, \theta = R/N)$, but a uniform prior on R does not converge to the commonly accepted (both Jeffreys and reference) continuous objective prior $\mathrm{Be}(\theta \mid \frac{1}{2}, \frac{1}{2})$, thus leading to mutually inconsistent inferences on $\theta = R/N$, even for very large N values.

  - As these examples suggest, if the θ's in $p(\boldsymbol{z} \mid \theta)$ are not just labels, but meaningful quantitative values, some structure may be needed in the reference prior (thus restricting the class $\mathcal{P}$ of permissible priors) to incorporate this assumed knowledge.

- *Structured reference priors for discrete parameters*

  - Embed the original model $p(\boldsymbol{z} \,|\, \theta)$, θ discrete, into a model $p_e(\boldsymbol{z} \,|\, \omega)$ with a continuous parameter ω, apply standard reference prior theory (Bernardo 1979, 2005, Berger, Bernardo and Sun, 2008a) to obtain $\pi_e(\omega)$, and appropriately discretize $\pi_e(\omega)$ to get $\pi(\theta)$.

  - No single embedding methodology seems to be always successful. The choice of embedding may be the discrete analog of the need to choose a sequence of compact sets in continuous models. Possible embedding methodologies (Berger, Bernardo and Sun, 2008b) include:

    (i) Treat the original parameter as continuous, after appropriate renormalization, if necessary.

    (ii) Derive the reference prior $\pi_e(\theta)$ for a continuous asymptotic sampling distribution $p_e(\hat{\theta} \,|\, \theta)$ of some consistent estimator of θ.

    (iii) Add a hierarchical structure $p(\theta \,|\, \omega)$ with continuous hyperparameter ω to the original model, derive the reference prior $\pi(\omega)$ for the integrated model $p(\boldsymbol{z} \,|\, \omega) = \sum_\theta p(\boldsymbol{z} \,|\, \theta) p(\theta \,|\, \omega)$, and use the integrated prior $\pi(\theta) = \int_\Omega p(\theta \,|\, \omega) \pi(\omega) \, \mathrm{d}\omega$.

2. Sampling from Finite Populations

- *Random sampling without replacement from a finite population*

  - Finite population of size N with R *conforming* (+) elements, where $0 \leq R \leq N$. The propability that a random sample of size n contains r conforming elements, is

    $$\Pr(r \mid n, R, N) = \text{Hy}(r \mid n, R, N) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}$$

    if $r = 0, \ldots, \min\{n, R\}$, and zero otherwise.

- *Bayes and Laplace uniform prior*

  - Uniform prior $\pi_u(R) = (N+1)^{-1}, \quad R = 0, \ldots, N.$

    Posterior $\pi_u(R \mid r, n, N) = \dfrac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N+1}{n+1}}, \quad R = r, \ldots, N - n + r.$

    In particular, $\Pr(\text{All} + \mid n, N) = \pi_u(R = N \mid r = n, N) = \frac{n+1}{N+1}$.

- *Laplace law of succession*

  - Probability that an element randomly selected among remaining unobserved $N - n$ elements is conforming is (Laplace, 1774)

    $$\Pr(+\,|\,r, n, N) = \sum_{R=r}^{N-n+r} \frac{R-r}{N-n} \Pr(R\,|\,r, N, n) = \frac{r+1}{n+2}\,.$$

    In particular, for $r = n$,

    $$\pi_u(E_n) = \pi_u(+\,|\,r = n, N) = \frac{n+1}{n+2}\,.$$

- *Succession and 'natural' induction*

  - With the uniform prior, if an event has been observed n uninterrupted times in a population of size N, it is very likely, $(n+1)/(n+2)$, that it will be observed again next time, but quite unlikely, $(n+1)/(N+1)$, that it will *always* be observed ('natural' induction).

  - For $\Pr(R = N\,|\,r = n, N)$ to be large with large $N >> n$, a different type of prior for R is needed. Jeffreys (1961) proposed priors of the form $\Pr(R = 0) = \Pr(R = N) = k$, $\frac{1}{3} \le k \le \frac{1}{2}$, with the remaining $1 - 2k$ uniformly distributed among the remaining $N - 1$ values of R.

3 Structured Reference Prior

- *Hypergeometric-binomial hierarchical model*

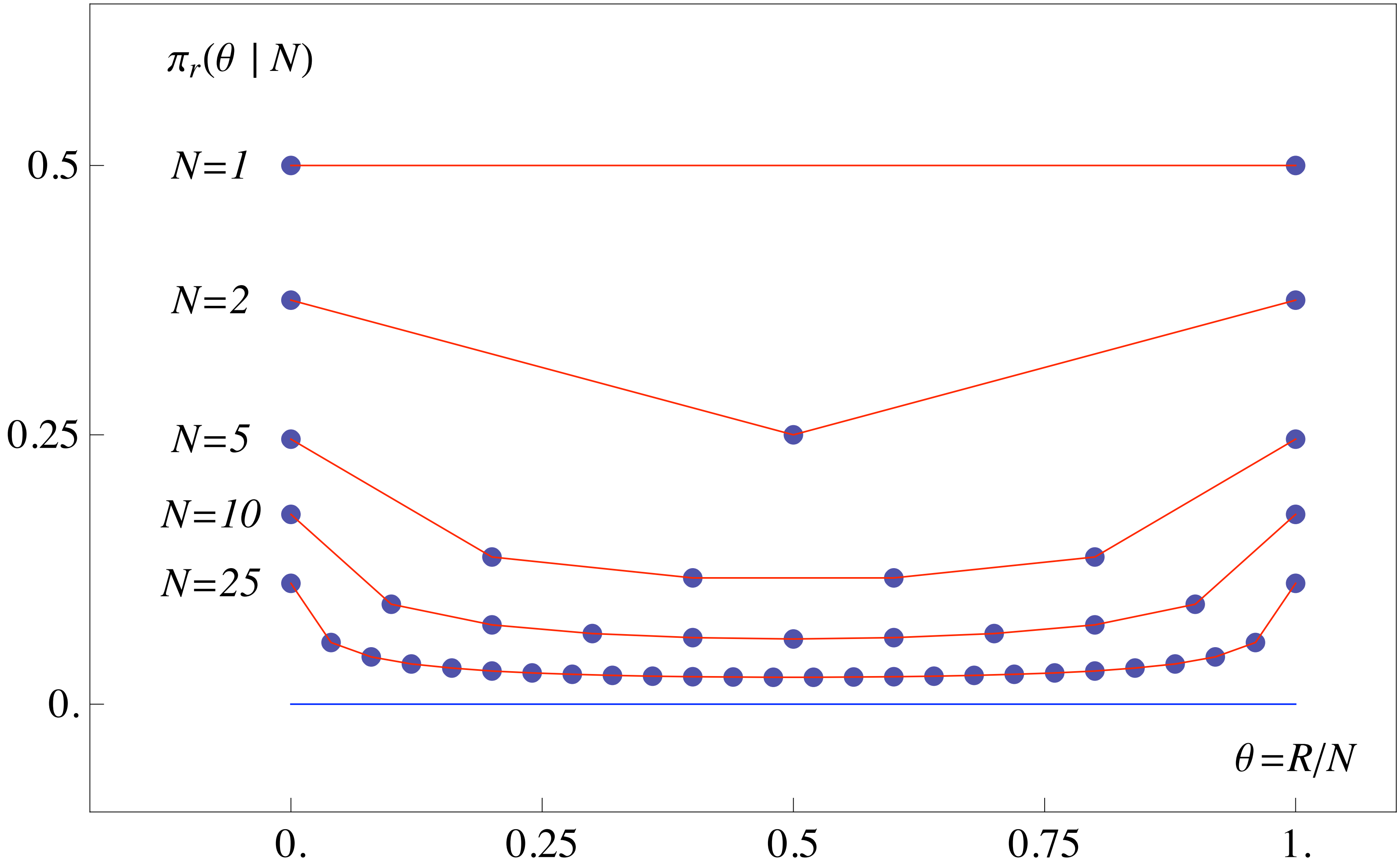
  - Consider the model $\text{Hy}(r \,|\, n, R, N)$ and assume that the R conforming items are a random sample from a binomial population with parameter p.

$$\begin{cases} \Pr(r \,|\, R, N, n) & = & \text{Hy}(r \,|\, n, R, N), \\ \Pr(R \,|\, N, p) & = & \text{Bi}(R \,|\, N, p). \end{cases}$$

  - A prior $\pi(p)$ must be chosen for the hyperparameter p. The appropriate reference prior (Bernardo and Smith, 1994, p. 339) is that which corresponds to the continuous parameter model obtained by eliminating R,
$\Pr(r \,|\, n, N, p) = \sum_{R=0}^{N} \text{Hy}(r \,|\, n, R, N)\, \text{Bi}(R \,|\, N, p) = \text{Bi}(r \,|\, n, p)$,
and this is Jeffreys prior $\pi(p) = \text{Be}(p \,|\, \frac{1}{2}, \frac{1}{2})$.

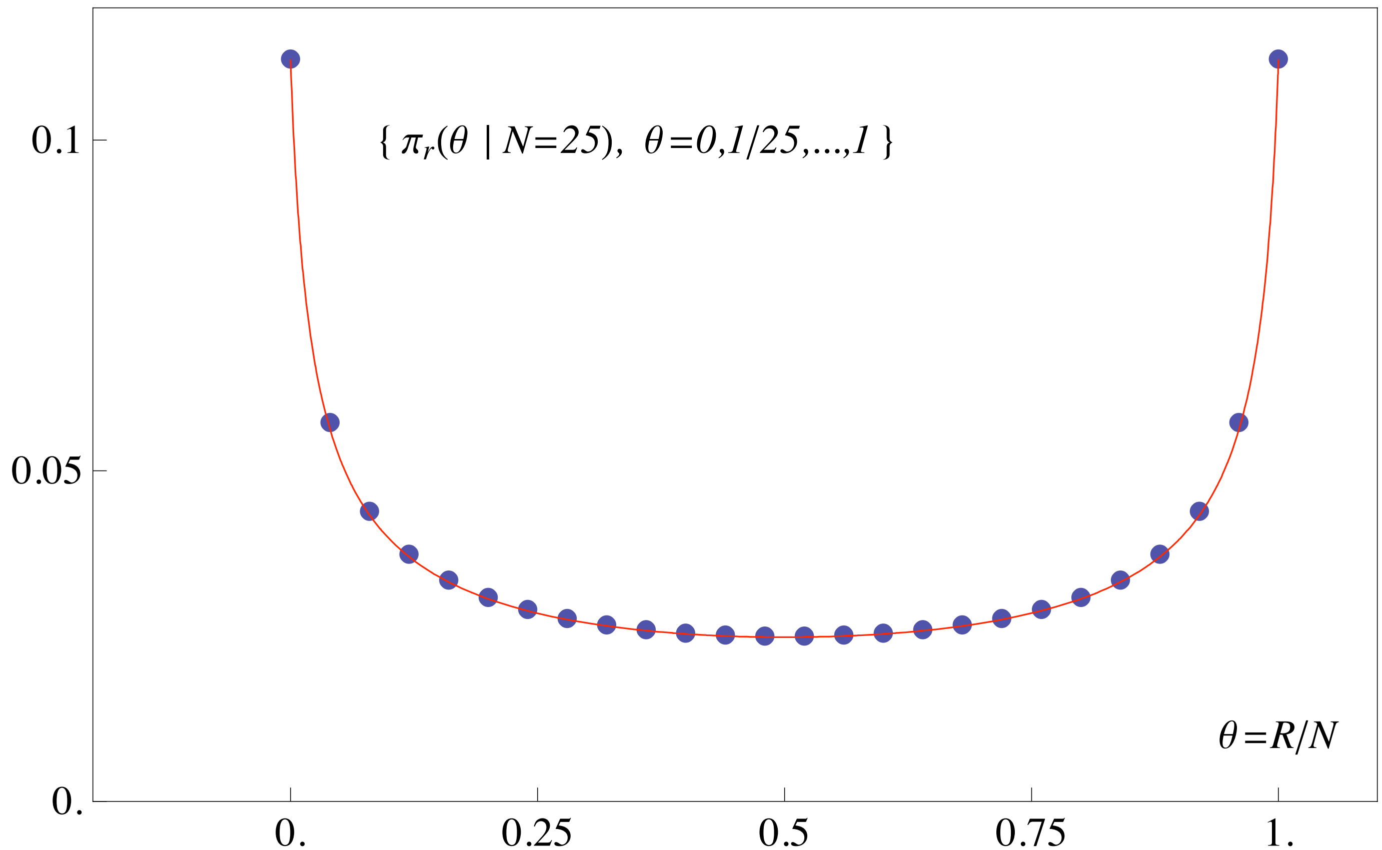
  - Hence, the corresponding structured reference prior for R, is
$\pi_r(R \,|\, N) = \int_0^1 \text{Bi}(R \,|\, N, p)\, \text{Be}(p \,|\, \frac{1}{2}, \frac{1}{2})\, \mathrm{d}p = \frac{1}{\pi} \dfrac{\Gamma(R+\frac{1}{2})\,\Gamma(N-R+\frac{1}{2})}{\Gamma(R+1)\,\Gamma(N-R+1)}$.

- *Structured priors for the hypergeometric model*



Reference prior probabilities $\pi_r(\theta = R/N \mid N)$, for several N values, $\theta \in \{0, 1/N, \ldots, (N-1)/N, 1\}$.

- *Large population size approximation*



□ For large N, using Stirling for the Gamma functions,

$$\pi_r(\theta \mid N) \approx \frac{1}{N+\frac{2}{\pi}} \mathrm{Be}\left(\frac{N\,\theta+\frac{1}{\pi}}{N+\frac{2}{\pi}} \mid \frac{1}{2}\,,\frac{1}{2}\right), \quad \theta = 0, 1/N, \ldots, 1,$$

which converges to the reference prior $\pi(\theta) = \mathrm{Be}(\theta \mid \frac{1}{2}, \frac{1}{2})$ as $N \to \infty$.

- *Reference prior predictive distribution of r*

  - The reference prior predictive distribution of the number r of conforming items in a random sample of size n is,

    $$\sum_{R=0}^{N} \mathrm{Hy}(r \,|\, R, N, n)\, \pi_r(R \,|\, N) = \frac{1}{\pi}\, \frac{\Gamma(r+\frac{1}{2}\,\Gamma(n-r+\frac{1}{2})}{\Gamma(r+1)\,\Gamma(n-r+1)} = \pi_r(r \,|\, n).$$

  - Notice that, as in the case of the uniform prior, the reference prior predictive distribution of r given n, has precisely the same mathematical form as the reference prior of R given N, $\pi_r(R \,|\, N)$.

- *Reference posterior distribution of R*

  - The reference posterior for R has the analytical form

    $$\pi_r(R \,|\, r, n, N) = c(r, N, n)\, \frac{\Gamma(R+\frac{1}{2})\,\Gamma(N-R+\frac{1}{2})}{\Gamma(R-r+1)\,\Gamma(N-R-(n-r)+1)}\,,$$

    $$c(r, N, n) = \frac{\Gamma(n+1)\,\Gamma(N-n+1)}{\Gamma(N+1)\,\Gamma(r+\frac{1}{2})\,\Gamma(n-r+\frac{1}{2})}\,.$$

    for $R \in \{r, r+1, \ldots, N-n+r\}$, and zero otherwise

- *Probability of all elements conforming*

  - In particular, for $R = N$ and $r = n$,

    $$\pi_r(\text{All} + \,|\, n, N) = \frac{\Gamma(N+1/2)}{\Gamma(N+1)} \frac{\Gamma(n+1)}{\Gamma(n+1/2)} \approx \sqrt{\frac{n}{N}}\;.$$

- *A new law of succesion*

  - Reference probability that a new element is conforming is

    $$\pi_r(+ \,|\, r, n, N) = \textstyle\sum_{R=r}^{N-n+r} \frac{R-r}{N-n}\, \pi_r(R \,|\, r, N, n) = \frac{r+1/2}{n+1}\;.$$

    In particular, for $r = n$,

    $$\pi_u(E_n) = \pi_u(+ \,|\, r = n, N) = \frac{n+1/2}{n+1}\;.$$

    Converges faster to one than Laplace as n increases. For $n = 1$ this yields $3/4$ rather than Lapace $2/3$.

  - As with the uniform prior, under the structured reference prior, if an event has been observed n uninterrupted times in a population of size N, it is very likely, $(n + 1/2)/(n + 1)$, that it will be observed next time, but quite unlikely, about $\sqrt{n/N}$, that it will *always* be observed.

4 Natural Induction

Testing the Precise Hypothesis that R=N

- *The need for a mixture prior*
  - Given a model $p(\boldsymbol{z} \mid \phi)$, derivation of a posterior probability $\Pr(H_0 \mid \boldsymbol{z})$ for a precise hypothesis $H_0 = \{\phi = \phi_0\}$ typically requires the use of a mixture prior which assigns assigns a lump of probability to $\{\phi = \phi\}$.
  - This may be obtained from standard use of reference analysis, if the parameter of interest is chosen to be whether or not $\{\phi = \phi_0\}$, rather than the actual value of ϕ. The result may be seen as a reformulation of the use of a Bayes factor to test the hypothesis that $\{\phi = \phi_0\}$ versus the alternative $\{\phi \neq \phi_0\}$.
  - In finite population sampling, the name 'natural' induction is often associated to testing whether or not all the elements in the popupation share a given characteristic, *i.e.*, testing the hypothesis that $R = N$ versus the alternative $R \neq N$.

- *Reference prior for testing $H_0 = \{R = N\}$*

  - In the model $\text{Hy}(r \,|\, n, R, N)$ let the quantity of interest be
  $$\phi = \begin{cases} \phi_0 & \text{if } R = N \text{ (All +)} \\ \phi_1 & \text{if } 0 \le R < N, \end{cases}$$
  and the nuisance parameter $\lambda = \begin{cases} \lambda_0 & \text{if } R = N \text{ (All +)} \\ R & \text{if } 0 \le R < N. \end{cases}$

  - Trivially, $\pi(\lambda = \lambda_0 \,|\, \phi = \phi_0, N) = 1$. Moreover, the sampling distribution of r given $\phi = \phi_1$ is $\text{Hy}(r \,|\, n, R, N-1)$ and, therefore, $\pi(\lambda \,|\, \phi = \phi_1, N) = \pi_r(R \,|\, N-1)$. Since ϕ has only two possible values, $\pi(\phi = \phi_0) = \pi(\phi = \phi_1) = 1/2$.

  - Hence, the joint reference prior $\pi_0(R \,|\, N)$ of the unknown parameter R when ϕ is the quantity of interest is
  $$\pi(\lambda \,|\, \phi, N)\, \pi(\phi) = \begin{cases} \frac{1}{2} & \text{if } R = N \\ \frac{1}{2}\, \frac{1}{\pi}\, \frac{\Gamma(R+1/2)\, \Gamma(N-1-R+1/2)}{\Gamma(R+1)\, \Gamma(N-1-R+1)} & \text{if } R \ne N. \end{cases}$$

- *Reference posterior probability* $\Pr(H_0 \,|\, n, N)$

  - Using Bayes theorem $\pi_0(\text{All} + \,|\, n, N)$ is given by

$$\pi_0(\phi = \phi_0 \,|\, r = n, N) = \frac{\frac{1}{2}\Pr(r{=}n \,|\, \phi{=}\phi_0{,}N)}{\frac{1}{2}\Pr(r{=}n \,|\, \phi{=}\phi_0{,}N) + \frac{1}{2}\Pr(r{=}n \,|\, \phi{=}\phi_1{,}N)},$$

  - But $\Pr(r = n \,|\, \phi = \phi_0, N) = \begin{cases} 1 & \text{if } r = n, \\ 0 & \text{if } 0 \leq r < n, \end{cases}$

    and $\Pr(r = n \,|\, \phi = \phi_1, N) = \sum_{R=n}^{N-1} \text{Hy}(n \,|\, n, R, N)\; \pi_r(R \,|\, N-1)$

  - Substitution and simplification yields

$$\pi_0(\text{All} + \,|\, n, N) = \frac{1}{1 + \frac{1}{\sqrt{\pi}}\,\frac{N-n}{N}\,\frac{\Gamma(n+1/2)}{\Gamma(n+1)}}$$

    Using Stirling, $\Gamma(n + 1/2)/\Gamma(n + 1) \approx 1/\sqrt{n + 1}$.

    If $N >> n$, $(N - n)/N \approx 1$.

    Thus, $\pi_0(\text{All} + \,|\, n, N) \approx \sqrt{\pi(n + 1)}/(1 + \sqrt{\pi(n + 1)}\,)$.

- *Continuous approximation Bayes factor approach*

  - Conventional testing of $H_0 = \{p = 1\}$ in a Binomial $\mathrm{Bi}(r \,|\, n, p)$ model uses the mixture prior

    $$\Pr(p) = \begin{cases} \frac{1}{2} & p = 1, \\ \frac{1}{2}\,\mathrm{Be}(p \,|\, \frac{1}{2}, \frac{1}{2}) & p \neq 1, \end{cases}$$

    and, given $r = n$, yields

    $$\Pr(H_0 \,|\, \boldsymbol{z}) = \frac{1}{1+\mathrm{BF}(H_0,n)} = \frac{1}{1+\frac{1}{\sqrt{\pi}}\frac{\Gamma(n+\frac{1}{2})}{\Gamma(n+1)}}$$

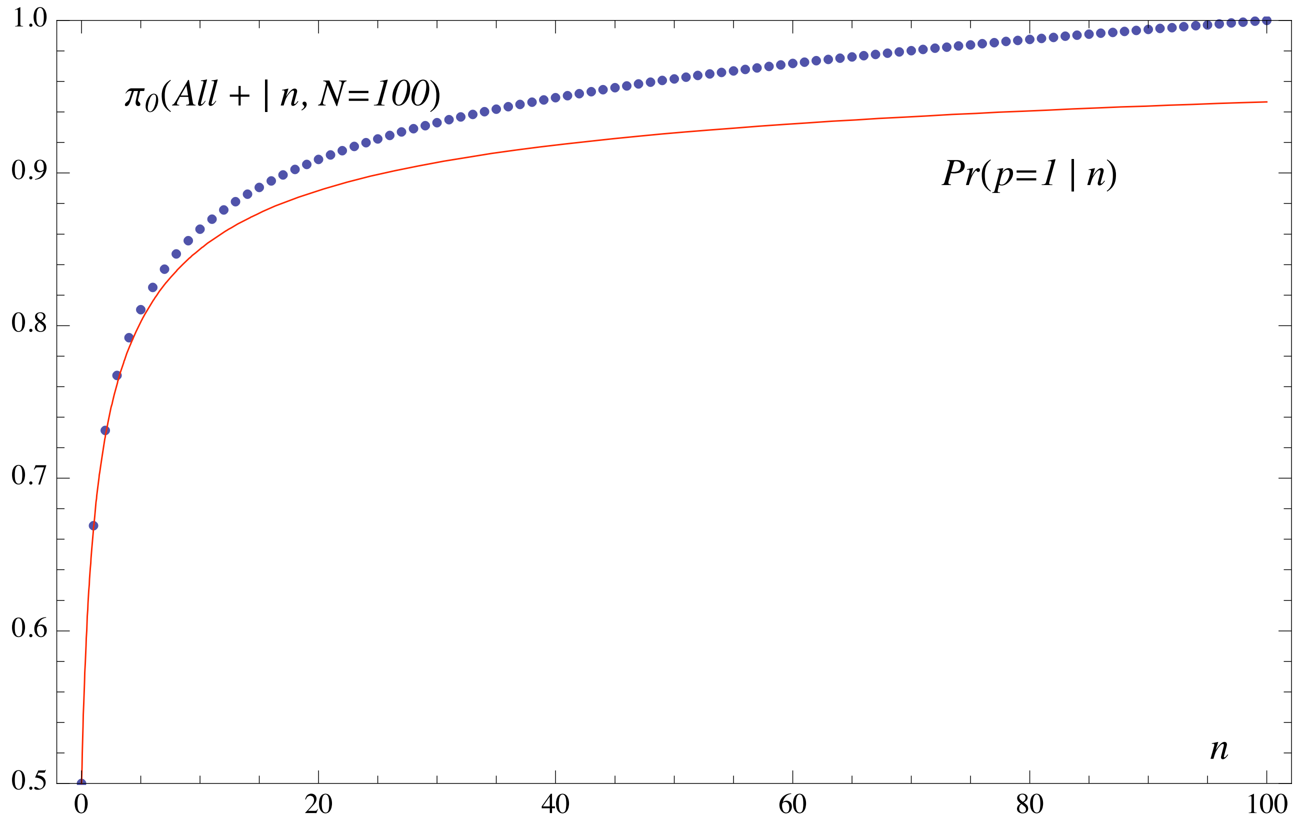
  - Except for the population size corrrection factor $(N - n)/N$, this is the reference posterior probability $\pi_0(\mathrm{All}+ \,|\, n, N)$. Thus, the use of a (conditional) structured reference prior produces results compatible with the accepted solution for the limiting continuous case.

  - With a uniform conditional prior the result (Bernardo, 1985) is

    $$\pi_1(\mathrm{All}+ \,|\, n, N) = (1 + \frac{1}{n+1}\,(1 - \frac{n}{N}))^{-1},$$

    which is *not* compatible with the limiting continuous result.

- *Reference posterior probability and continuous Bayes factor*



$$\pi_0(\mathrm{All}+\,|\,n,N)=\frac{1}{1+\frac{1}{\sqrt{\pi}}\,\frac{N-n}{N}\,\frac{\Gamma(n+1/2)}{\Gamma(n+1)}}$$

$$\Pr(p=1\,|\,n)=\frac{1}{1+\frac{1}{\sqrt{\pi}}\frac{\Gamma(n+\frac{1}{2})}{\Gamma(n+1)}}$$

- *Example: Galapagos Islands*

  - Charles Darwin research station, Galapagos Islands, Pacific Ocean.
  - A zoologist observes and marks 55 galapagos in a particular island, all of which present a particular shell modification. What is the *probability* that all galapagos in theat island have the reported modification?

    Assume sample is random, and $N \in [150, 250]$.
  - A conditional uniform prior yields the range $[0.986, 0.989]$. The structured reference condional prior gives the considerably lower range

    $\pi_0(\text{All} + \,|\, n = 55, N \in [150, 250]) \in [0.944, 0.954]$,

    still higher than the continuous the Bayes factor approximation 0.929.
  - Besides, the predictive probability that the first new unmarked galapago to be observed in that also presents a modified shell is

    $\pi_r(+ \,|\, r = n = 55) = 0.991$.

- *Other Examples*

  - Quality assurance problems. Pharmacology. Physical Sciences.

References

Available on line at **www.uv.es/bernardo**

Berger, J. O., Bernardo, J. M. and Sun, D. (2008a). The formal definition of reference priors. *Annals of Statistics*, **36** (in press).

Berger, J. O., Bernardo, J. M. and Sun, D. (2008b). Reference priors for discrete parameter spaces. *Tech. Rep.*, Universidad de Valencia, Spain.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113—147, (with discussion).

Bernardo, J. M. (1985). On a famous problem of induction. *Trabajos Estadist.* **36**, 24–30.

Bernardo, J. M. (2005). Reference analysis. *Handbook of Statistics* **25** (D. K. Dey and C. R. Rao eds.). Amsterdam: Elsevier, 17–90.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley, (2nd. edition in preparation).

Jeffreys, H. (1961). *Theory of Probablity*, 3rd ed. Oxford: University Press.

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les événements. *Oeuvres Complètes* **8**, 27–68. Paris: Gauthier-Villars, 1891.

Valencia Mailing List

- The Valencia Mailing List contains about 2,000 entries of people interested in Bayesian Statistics. It sends information about the Valencia Meetings and other material of interest to the Bayesian community.

- Last Proceedings volume:

  Bernardo, J.M., Bayarri, M.J., Berger, J.O. Dawid, A.P. Heckerman, D., Smith, A.F.M. and West, M. (eds.) (2007).
  Bayesian Statistics 8. Oxford, UK: University Press.

- Next Conference:

  9th Valencia International Meeting on Bayesian Statistics
  ISBA World Meeting 2010
  State of Valencia (Spain), June 2010

- If you do not belong to the list and want to be included, please send your data to **<valenciameeting@uv.es>**

Gracias por vuestra atención!

jose.m.bernardo@uv.es

www.uv.es/bernardo

valenciameeting@uv.es

www.uv.es/valenciameeting