Bayes and Discovery: Objective Bayesian Hypothesis Testing

José M. Bernardo

Universitat de València, Spain

jose.m.bernardo@uv.es

PHYSTAT 2011

CERN (Geneva, Switzerland), January 17th, 2011

Summary

Bayesian Inference Summaries

- Assume data $\boldsymbol{z}$ have been generated as one random observation form $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, where $\boldsymbol{\theta}$ is the vector of interest and $\boldsymbol{\lambda}$ a nuisance parameter vector.
- Let $p(\boldsymbol{\theta},\boldsymbol{\lambda}) = p(\boldsymbol{\lambda}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta})$ be the assumed joint prior.
- Given data $\boldsymbol{z}$ and assuming model $\mathcal{M}_{\boldsymbol{z}}$, the complete solution to all inference questions about $\boldsymbol{\theta}$ is contained in the marginal posterior $p(\boldsymbol{\theta}\,|\,\boldsymbol{z})$, derived by standard use of probability theory.
- Appreciation of $p(\boldsymbol{\theta}\,|\,\boldsymbol{z})$ may be enhanced by providing both point and region estimates of the vector of interest $\boldsymbol{\theta}$, and by declaring whether or not some context-suggested specific value $\boldsymbol{\theta}_0$ (or maybe a set of values $\boldsymbol{\Theta}_0$), is (are) compatible with the observed data $\boldsymbol{z}$. These elaborations provide useful (and often required) summaries of $p(\boldsymbol{\theta}\,|\,\boldsymbol{z})$.

Decision-theoretic structure

- All these summaries may be framed as different decision problems which use precisely the same loss function $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ describing, as a function of the (unknown) $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ values which have generated the available data $\boldsymbol{z}$, the loss to be suffered if, working with model $\mathcal{M}_{\boldsymbol{z}}$, the value $\boldsymbol{\theta}_0$ were used as a proxy for the unknown value of $\boldsymbol{\theta}$.

- The results dramatically depend on the choices made for both the prior and the loss functions but (given $\boldsymbol{z}$) only depend on those through the expected loss, $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, p(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z})\, d\boldsymbol{\theta} d\boldsymbol{\lambda}$.

- As a function of $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ is a measure of the unacceptability of all possible values of the vector of interest. This provides a dual, complementary information on all $\boldsymbol{\theta}$ values (on a loss scale) to that provided by the posterior $p(\boldsymbol{\theta} \,|\, \boldsymbol{z})$ (on a probability scale).

□ Point estimation

To choose a point estimate for $\boldsymbol{\theta}$ is a decision problem where the action space is the class $\boldsymbol{\Theta}$ of all possible $\boldsymbol{\theta}$ values.

Definition 1 *The Bayes estimator* $\boldsymbol{\theta}^*(\boldsymbol{z}) = \arg\inf_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}} \overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ *is that which minimizes the posterior expected loss.*

- Conventional examples include the ubiquitous quadratic loss $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = (\boldsymbol{\theta}_0 - \boldsymbol{\theta})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta})$, which yields the posterior mean as the Bayes estimator, and the zero-one loss on a neighborhood of the true value, which yields the posterior mode a a limiting result.
- Bayes estimators with conventional loss functions are typically not invariant under one to one transformations. Thus, the Bayes estimator under quadratic loss of a variance is not the square of the Bayes estimator of the standard deviation. This is rather difficult to explain when one merely wishes to report an estimate of some quantity of interest.

Region estimation

Bayesian region estimation is achieved by quoting posterior credible regions. To choose a q-credible region is a decision problem where the action space is the class of subsets of $\boldsymbol{\Theta}$ with posterior probability q.

Definition 2 *(Bernardo, 2005). A* Bayes q-credible region $\boldsymbol{\Theta}^*_q(\boldsymbol{z})$ *is a q-credible region where any value within the region has a smaller posterior expected loss than any value outside the region:*
$\forall \boldsymbol{\theta}_i \in \boldsymbol{\Theta}^*_q(\boldsymbol{z}),\ \forall \boldsymbol{\theta}_j \notin \boldsymbol{\Theta}^*_q(\boldsymbol{z}),\quad \bar{\ell}(\boldsymbol{\theta}_i \mid \boldsymbol{z}) \leq \bar{\ell}(\boldsymbol{\theta}_j \mid \boldsymbol{z}).$

- The quadratic loss yields credible regions with those $\boldsymbol{\theta}$ values closest, in the Euclidean sense, to the posterior mean. A zero-one loss function leads to highest posterior density (HPD) credible regions.
- Conventional Bayes regions are typically not invariant: HPD regions in one parameterization will not transform to HPD regions in another.

☐ Precise hypothesis testing

• Consider a value $\boldsymbol{\theta}_0$ which deserves special consideration. Testing the hypothesis $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ is as a decision problem where the action space $\mathcal{A} = \{a_0, a_1\}$ contains only two elements: to accept (a_0) or to reject (a_1) the hypothesis H_0.

• Foundations require to specify the loss functions $\ell_h\{a_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ and $\ell_h\{a_1, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ measuring the consequences of accepting or rejecting H_0 as a function of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. The optimal action is to reject H_0 iif

$$\int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} [\ell_h\{a_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_h\{a_1, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}] \, p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{z}) \, d\boldsymbol{\theta} d\boldsymbol{\lambda} > 0.$$

• Hence, only $\Delta\ell_h\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \ell_h\{a_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_h\{a_1, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, which measures the conditional advantage of rejecting, must be specified.

- Without loss of generality, the function $\Delta\ell_h$ may be written as

$$\Delta\ell_h\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_0$$

where (precisely as in estimation), $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ describes, as a function of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, the non-negative loss to be suffered if $\boldsymbol{\theta}_0$ were used as a proxy for the (unknown) true value of $\boldsymbol{\theta}$.

- Since $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}_0, \boldsymbol{\lambda})\} = 0$, the constant $\ell_0 > 0$ measures the (context-dependent) positive advantage of accepting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ when it is true.

Definition 3 *(Bernardo and Rueda, 2002). The* Bayes test criterion *to decide on the compatibility of $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with available data $\boldsymbol{z}$ is to reject $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ if (and only if), $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) > \ell_0$, where ℓ_0 is a context dependent positive constant.*

- The compound case may be analyzed by separately considering each of the values which make part of the compound hypothesis to test.

• Using a zero-one loss function, so that the loss advantage of rejecting $\boldsymbol{\theta}_0$ is equal to one whenever $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and zero otherwise, leads to rejecting H_0 if (and only if) $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) < p_0$ for some context-dependent p_0. Use of this loss requires the prior probability $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0)$ to be strictly positive. If $\boldsymbol{\theta}$ is a continuous parameter this forces the use of a non-regular "sharp" prior, concentrating a positive probability mass at $\boldsymbol{\theta}_0$, the solution early advocated by Jeffreys.

This formulation (i) implies the use of radically different priors for hypothesis testing than those used for estimation, (ii) precludes the use of conventional, often improper, 'noninformative" priors, and (iii) may lead to the difficulties associated to Jeffreys-Lindley paradox.

• The quadratic loss function leads to rejecting a $\boldsymbol{\theta}_0$ value whenever its Euclidean distance to $\mathrm{E}[\boldsymbol{\theta} \,|\, \boldsymbol{z}]$, the posterior expectation of $\boldsymbol{\theta}$, is sufficiently large.

- The use of continuous loss functions (such as the quadratic loss) permits the use in hypothesis testing of precisely the same priors that are used in estimation.
- With conventional loss functions the Bayes test criterion is typically not invariant under one-to-one transformations. Thus, if $\boldsymbol{\phi}(\boldsymbol{\theta})$ is a one-to-one transformation of $\boldsymbol{\theta}$, rejecting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ does not generally imply rejecting $\boldsymbol{\phi}(\boldsymbol{\theta}) = \boldsymbol{\phi}(\boldsymbol{\theta}_0)$, a rather unpalatable situation.

- The threshold constant ℓ_0, which controls whether or not an expected loss is too large, is part of the specification of the decision problem, and should be context-dependent. However a judicious choice of the loss function leads to calibrated expected losses, where the relevant threshold constant has an immediate, operational interpretation.

Loss Functions

- A dissimilarity measure $\delta\{p_{\boldsymbol{z}}, q_{\boldsymbol{z}}\}$ between two probability densities $p_{\boldsymbol{z}}$ and $q_{\boldsymbol{z}}$ for a random vector $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ should be

  (i) non-negative, and zero if (and only if) $p_{\boldsymbol{z}} = q_{\boldsymbol{z}}$ a.e.,

  (ii) invariant under one-to-one transformations of $\boldsymbol{z}$,

  (iii) symmetric, so that $\delta\{p_{\boldsymbol{z}}, q_{\boldsymbol{z}}\} = \delta\{q_{\boldsymbol{z}}, p_{\boldsymbol{z}}\}$,

  (iv) defined for densities with strictly nested supports.

Definition 4 *The* intrinsic discrepancy $\delta\{p_1, p_2\}$ *is*

$$\delta\{p_1, p_2\} = \min\left[\,\kappa\{p_1 \,|\, p_2\},\ \kappa\{p_2 \,|\, p_1\}\,\right]$$

where $\kappa\{p_j \,|\, p_i\} = \int_{\boldsymbol{\mathcal{Z}}_i} p_i(\boldsymbol{z}) \log[p_i(\boldsymbol{z})/p_j(\boldsymbol{z})]\, d\boldsymbol{z}$ *is the* (KL) *divergence of* p_j *from* p_i*. The intrinsic discrepancy between* p *and a family* $\mathcal{F} = \{q_i, i \in I\}$ *is the intrinsic discrepancy between* p *and the closest element in* $\mathcal{F}$*,* $\delta\{p, \mathcal{F}\} = \inf_{q,\in\mathcal{F}} \delta\{p, q\}$.

The intrinsic loss function

Definition 5 *Consider* $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda}), \boldsymbol{z}\in\boldsymbol{\mathcal{Z}}, \boldsymbol{\theta}\in\boldsymbol{\Theta}, \boldsymbol{\lambda}\in\boldsymbol{\Lambda}\}$. *The* intrinsic loss *of using* $\boldsymbol{\theta}_0$ *as a proxy for* $\boldsymbol{\theta}$ *is the intrinsic discrepancy between the true model and the class of models with* $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, $\mathcal{M}_0 = \{p(\boldsymbol{z}\,|\,\boldsymbol{\theta}_0,\boldsymbol{\lambda}_0), \boldsymbol{z}\in\boldsymbol{\mathcal{Z}}, \boldsymbol{\lambda}_0\in\boldsymbol{\Lambda}\}$,

$$\ell_\delta\{\boldsymbol{\theta}_0,(\boldsymbol{\theta},\boldsymbol{\lambda})\,|\,\mathcal{M}_{\boldsymbol{z}}\} = \inf_{\boldsymbol{\lambda}_0\in\boldsymbol{\Lambda}} \delta\{p_{\boldsymbol{z}}(\cdot\,|\,\boldsymbol{\theta},\boldsymbol{\lambda}), p_{\boldsymbol{z}}(\cdot\,|\,\boldsymbol{\theta}_0,\boldsymbol{\lambda}_0)\}.$$

☐ Invariance

- For any one-to-one reparameterization $\boldsymbol{\phi}=\boldsymbol{\phi}(\boldsymbol{\theta})$ and $\boldsymbol{\psi}=\boldsymbol{\psi}(\boldsymbol{\theta},\boldsymbol{\lambda})$,

$$\ell_\delta\{\boldsymbol{\theta}_0,(\boldsymbol{\theta},\boldsymbol{\lambda})\,|\,\mathcal{M}_{\boldsymbol{z}}\} = \ell_\delta\{\boldsymbol{\phi}_0,(\boldsymbol{\phi},\boldsymbol{\psi})\,|\,\mathcal{M}_{\boldsymbol{z}}\}.$$

This yields invariant Bayes point and region estimators, and invariant Bayes hypothesis testing procedures.

□ Reduction to sufficient statistics

• If $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{z})$ is a sufficient statistic for model $\mathcal{M}_{\boldsymbol{z}}$, one may also work with marginal model $\mathcal{M}_{\boldsymbol{t}} = \{p(\boldsymbol{t} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{t} \in \boldsymbol{\mathcal{T}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ since

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{t}}\}.$$

□ Additivity

• If data consist of a random sample $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from some model $\mathcal{M}_{\boldsymbol{x}}$, so that $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{X}}^n$, and $p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \prod_{i=1}^n p(\boldsymbol{x}_i \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = n\, \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{x}}\}.$$

This "likelihood friendly" property considerably simplifies frequent computations.

Objective Bayesian Methods

- The methods described so far may be used with any prior. However, an "objective" procedure, where the prior function is intended to describe a situation where there is no relevant information about the quantity of interest, is often required.
- Objectivity is a very emotionally charged word, and it should be explicitly qualified. No statistical analysis is really objective (both the experimental design and the model have strong subjective inputs). However, frequentist procedures are branded as "objective" just because their conclusions are only conditional on the model assumed and the data obtained. Bayesian methods where the prior function is derived from the assumed model are objective is this limited, but precise sense.

☐ Development of objective priors

- Vast literature devoted to the formulation of objective priors.
- Reference analysis (Bernardo, 1979, 2005a, 2011; Berger and Bernardo, 1989, 1992a,b,c; Berger, Bernardo and Sun, 2009, 2011) is possibly the better accepted approach.
- Very general, easily computable one-parameter result:

Theorem 1 *Let* $\boldsymbol{z}^{(k)} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$ *denote* k *conditionally independent observations from* $\mathcal{M}_{\boldsymbol{z}}$*. For sufficiently large* k

$$\pi_k(\theta) \propto \exp\left\{\mathrm{E}_{\boldsymbol{z}^{(k)}\,|\,\theta}[\,\log p_h(\theta \,|\, \boldsymbol{z}^{(k)})]\right\}$$

where $p_h(\theta \,|\, \boldsymbol{z}^{(k)}) \propto \prod_{i=1}^{k} p(\boldsymbol{z}_i \,|\, \theta)\, h(\theta)$ *is the posterior which corresponds to any arbitrarily chosen strictly positive prior function* $h(\theta)$ *which makes the posterior proper for any* $\boldsymbol{z}^{(k)}$*.*

□ Approximate reference priors

• Reference priors are derived for an ordered parameterization. Given $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\omega}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ with m parameters, the reference prior with respect to $\boldsymbol{\phi}(\boldsymbol{\omega}) = \{\phi_1, \ldots, \phi_m\}$ is sequentially obtained as $\pi(\boldsymbol{\phi}) = \pi(\phi_m \,|\, \phi_{m-1}, \ldots, \phi_1) \times \cdots \times \pi(\phi_2 \,|\, \phi_1)\, \pi(\phi_1)$.

• One is often simultaneously interested in several functions of the parameters. Given $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\omega}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\omega} \in \boldsymbol{\Omega} \subset \Re^m\}$ with m parameters, consider a set $\boldsymbol{\theta}(\boldsymbol{\omega}) = \{\theta_1(\boldsymbol{\omega}), \ldots, \theta_r(\boldsymbol{\omega})\}$ of $r > 1$ functions of interest; Berger, Bernardo and Sun (work in progress) suggest a procedure to select a joint prior $\pi_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ whose corresponding marginal posteriors $\{\pi_{\boldsymbol{\theta}}(\theta_i \,|\, \boldsymbol{z})\}_{i=1}^r$ will be close, for all possible data sets $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$, to the set of reference posteriors $\{\pi(\theta_i \,|\, \boldsymbol{z})\}_{i=1}^r$ yielded by the set of reference priors $\{\pi_{\theta_i}(\boldsymbol{\omega})\}_{i=1}^r$ derived under the assumption that each of the θ_i's is of interest.

Definition 6 *Consider model* $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \mid \boldsymbol{\omega}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ *and* $r > 1$ *functions of interest,* $\{\theta_1(\boldsymbol{\omega}), \ldots, \theta_r(\boldsymbol{\omega})\}$*. Let* $\{\pi_{\theta_i}(\boldsymbol{\omega})\}_{i=1}^r$ *be the relevant reference priors, and* $\{\pi_{\theta_i}(\boldsymbol{z})\}_{i=1}^r$ *and* $\{\pi(\theta_i \mid \boldsymbol{z})\}_{i=1}^r$ *the corresponding prior predictives and marginal posteriors. Let* $\mathcal{F} = \{\pi(\boldsymbol{\omega} \mid \boldsymbol{a}), \boldsymbol{a} \in \mathcal{A}\}$ *be a family of prior functions. For each* $\boldsymbol{\omega} \in \boldsymbol{\Omega}$*, the best approximate joint reference prior within* $\mathcal{F}$ *is that which minimizes the average expected intrinsic loss*

$$d(\boldsymbol{a}) = \frac{1}{r} \sum_{i=1}^{r} \int_{\boldsymbol{\mathcal{Z}}} \delta\{\pi_{\theta_i}(\cdot \mid \boldsymbol{z}),\, p_{\theta_i}(\cdot \mid \boldsymbol{z}, \boldsymbol{a})\}\, \pi_{\theta_i}(\boldsymbol{z})\, d\boldsymbol{z}, \quad \boldsymbol{a} \in \mathcal{A}.$$

- Example. Use of the Dirichlet family in the m-multinomial model (with $r = m + 1$ cells) yields $\mathrm{Di}(\boldsymbol{\theta} \mid 1/r, \ldots, 1/r)$, with important applications to sparse multinomial data and contingency tables.

Integrated Reference Analysis

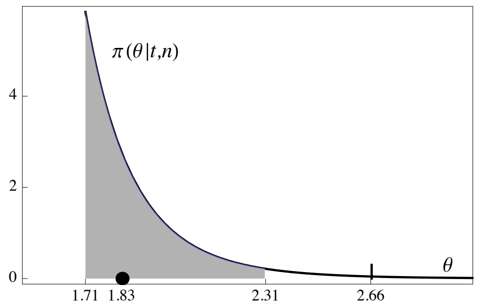
• We suggest a systematic use of the intrinsic loss function, and an appropriate joint reference prior, for an integrated objective Bayesian solution to both estimation and hypothesis testing in pure inference problems.

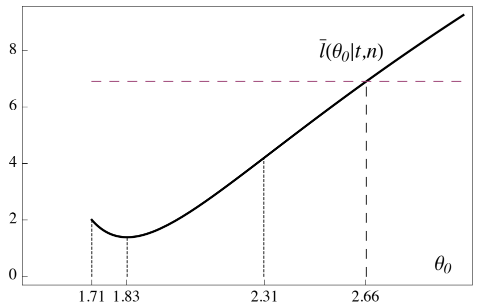
• We have stressed foundations-based decision theoretic arguments. Besides a large collection of detailed, non-trivial examples prove that the procedures advocated lead to attractive, often novel solutions. Details in Bernardo (2011), and references therein.

☐ Estimation of the normal variance

• The intrinsic (invariant) point estimator of the normal standard deviation is $\sigma^* \approx \frac{n}{n-1}\, s$. Hence, $\sigma^{2*} \approx \frac{n}{n-1}\frac{ns^2}{n-1}$, larger than both the mle s^2 and the unbiased estimator $ns^2/(n-1)$.

Uniform model $\mathrm{Un}(x \mid 0, \theta)$





$$\ell_\delta\{\theta_0, \theta \mid \mathcal{M}_{\boldsymbol{z}}) = n \begin{cases} \log(\theta_0/\theta), & \text{if } \theta_0 \geq \theta, \\ \log(\theta/\theta_0, & \text{if } \theta_0 \leq \theta. \end{cases}$$

$\pi(\theta) = \theta^{-1}, \quad \boldsymbol{z} = \{x_1, \ldots, x_n\},$

$t = \max\{x_1, \ldots, x_n\}, \quad \pi(\theta \mid \boldsymbol{z}) = n\, t^n \theta^{-(n+1)}$

The q-quantile is $\theta_q = t\,(1-q)^{-1/n}$;

Exact probability matching.

$\theta^* = t\, 2^{1/n}$ (posterior median)

$\mathrm{E}[\overline{\ell}_\delta(\theta_0 \mid t, n) \mid \theta] = (\theta/\theta_0)^n - n\log(\theta/\theta_0)$;
this is equal to 1 if $\theta = \theta_0$,
and increases with n otherwise.

- Simulation: $n = 10$ with $\theta = 2$ which yielded $t = 1.71$;
  $\theta^* = 1.83$, $\Pr[t < \theta < 2.31 \mid \boldsymbol{z}] = 0.95$, $\overline{\ell}_\delta(2.66 \mid \boldsymbol{z}) = \log 1000$.

Objective Bayesian Hypothesis Testing

- Assuming model $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, to test $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$, compute the expected reference intrinsic loss,

$$d(H_0 \,|\, \boldsymbol{z}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} \, \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z}) \, d\boldsymbol{\theta} d\boldsymbol{\lambda},$$

where $\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ is the intrinsic discrepancy between the true model and the family of models $\mathcal{M}_0 = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, , \boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}\}$ which satisfy H_0, and $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z})$ is the joint reference posterior when $\boldsymbol{\theta}$ is the vector of interest.

- Reject H_0 iff $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) > d_0$, where $d_0 > 0$ is context dependent.
- The function $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ is the intrinsic test statistic.
- Large values of d_0 correspond to situations with large advantages for accepting H_0 when it is true.

The choice of the threshold constant

- Under regularity conditions the intrinsic discrepancy reduces to

$$\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \int_{\mathcal{Z}} p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) \log \frac{p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})}{p(\boldsymbol{z} \mid \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0)} \, d\boldsymbol{z},$$

the minimum log-likelihood ratio against the null which may be expected under repeated sampling from the assumed model, and $d(H_0 \mid \boldsymbol{z})$ is just the posterior expectation (given the available data) of this quantity.

- The choice $d_0 = \log K$ therefore implies that H_0 is rejected when the average log-likelihood ratio against H_0 is expected to be larger than $\log K$.

- Simple choices of d_0 are $\{\log 10, \log 100, \log 1000\} \approx \{2.3, 4.6, 6.9\}$, which respectively suggest mild, moderate and strong evidence against H_0.

Interpretation

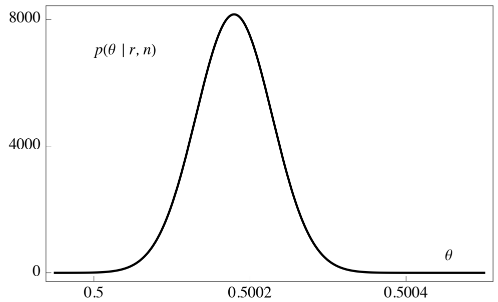
- As described above, the threshold d_0 has a simple operational interpretation in terms of acceptable average log-likelihood ratios against H_0. Of course, one may also simply quote $d(H_0 \,|\, \boldsymbol{z})$, and describe this as the posterior expectation (given the data) of the average (under sampling) log-likelihood ratio against H_0, without making any formal decision about accepting or rejecting H_0.

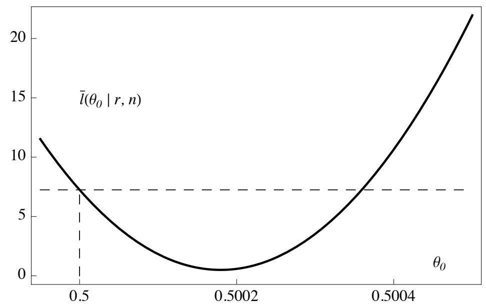
- The intrinsic test statistic $d(H_0 \,|\, \boldsymbol{z})$ is often a one-to-one transformation of conventional test statistics, but the ubiquitous $\alpha = 0.05$ frequentist choice then corresponds to $K \approx 11$, hardly strong evidence against H_0 (which explains the frequent false rejections found in the scientific literature).

Properties

- *Marginalization consistency*. Intrisic testing is consistent under reduction to sufficient statistics. Thus if a sufficient statistic $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{z})$ exists, testing a hypothesis using the full model $\mathcal{M}_{\boldsymbol{z}}$ is precisely equivalent to testing the hypothesis using the marginal model $\mathcal{M}_{\boldsymbol{t}}$ provided by the sampling distribution of $\boldsymbol{t}$.
- *Invariance under reparameterization*. Intrisic testing is an invariant procedure under reparameterization. Thus, for any one-to-one transformation $\boldsymbol{\phi}(\boldsymbol{\theta})$, the hyphothesis $\boldsymbol{\phi} = \boldsymbol{\phi}_0$ is accepted (rejected) if, and only if, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is accepted (rejected). This rather obvious coherency requirement is not, however, satisfied by many testing procedures (both frequentist and Bayesian)

☐ Extra Sensory Power (ESP) testing





Jahn, Dunne and Nelson (1987)
Binomial data. Test $H_0 \equiv \{\theta = 1/2\}$
with $n = 104,490,000$ and $r = 52,263,471$.
For any sensible continuous prior $p(\theta)$,
$p(\theta \mid \boldsymbol{z}) \approx \mathrm{N}(\theta \mid m_{\boldsymbol{z}}, s_{\boldsymbol{z}})$,
with $m_{\boldsymbol{z}} = (r + 1/2)/(n + 1) = 0.50018$,
$s_{\boldsymbol{z}} = [m_{\boldsymbol{z}}(1 - m_{\boldsymbol{z}})/(n+2)]^{1/2} = 0.000049$.
$d(H_0 \mid \boldsymbol{z}) \approx \frac{n}{2} \log[1 + \frac{1}{n}(1 + t_{\boldsymbol{z}}(\theta_0)^2)]$,
$t_{\boldsymbol{z}}(\theta_0) = (\theta_0 - m_{\boldsymbol{z}})/s_{\boldsymbol{z}}$, $t_{\boldsymbol{z}}(1/2) = 3.672$.
$d(H_0 \mid \boldsymbol{z}) = 7.24 = \log 1400$: Reject H_0

- Jeffreys-Lindley paradox: With any "sharp" prior, $\Pr[\theta = 1/2] = p_0$, $\Pr[\theta = 1/2 \mid \boldsymbol{z}] > p_0$ (Jefferys, 1990) suggesting data support H_0 !!!

☐ More sophisticated examples

- **Two sample problems**: **Equality of two normal means**.
  $d(H_0 \,|\, \boldsymbol{z}) \approx n \log[1 + \frac{1}{2n}(1 + t^2)], \quad t = \sqrt{n}(\bar{x} - \bar{y})/(s/\sqrt{2})$.
- **Trinomial data**: Testing for **Hardy-Weinberg equilibrium**.
  $d(H_0 \,|\, \boldsymbol{z}) \approx \int_{\mathcal{A}} \delta\{H_0, (\alpha_1, \alpha_2)\}\, \pi(\alpha_1, \alpha_2 \,|\, \boldsymbol{z}) d\alpha_1 d\alpha_2$,
  where $\delta\{H_0, (\alpha_1, \alpha_2)\} \approx n\, \theta(\alpha_1, \alpha_2)$,
  $\theta(\alpha_1, \alpha_2)$ is the KL distance of H_0 from $\mathrm{Tri}(r_1, r_2, r_3 \,|\, \alpha_1, \alpha_2)$ and
  $\pi(\alpha_1, \alpha_2 \,|\, \boldsymbol{z}) = \mathrm{Di}[\alpha_1, \alpha_2 \,|\, r_1 + 1/3, r_2 + 1/3, r_3 + 1/3]$.
- **Contingency tables**: Testing for **independence**.
  Data $\boldsymbol{z} = \{\{n_{11}, \ldots, n_{1b}\}, \ldots, \{n_{a1}, \ldots, n_{ab}\}\}$, $k = a \times b$,
  $d(H_0 \,|\, \boldsymbol{z}) \approx \int_{\boldsymbol{\Theta}} n\, \phi(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{z})\, d\boldsymbol{\theta}, \quad \phi(\boldsymbol{\theta}) = \sum_{i=1}^{a} \sum_{j=1}^{b} \theta_{ij} \log[\frac{\theta_{ij}}{\alpha_i\, \beta_j}]$,
  where $\alpha_i = \sum_{j=1}^{b} \theta_{ij}$ and $\beta_j = \sum_{i=1}^{a} \theta_{ij}$ are the marginals, and
  $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{z}) = \mathrm{Di}_{k-1}(\boldsymbol{\theta} \,|\, n_{11} + 1/k, \ldots, n_{ab} + 1/k)$.

Basic References

(In chronological order)

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

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200–207.

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

Berger, J. O. and Bernardo, J. M. (1992b). Reference priors in a variance components problem. *Bayesian Analysis is Statistics and Econometrics* (P. K. Goel and N. S. Yyengar, eds.) Berlin: Springer, 323–340.

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

Bernardo, J. M. (1997). Noninformative priors do not exist *J. Statist. Planning and Inference* **65**, 159–189 (with discussion).

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *Internat. Statist. Rev.* **70**, 351–372.

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

Bernardo, J. M. (2005b). Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test* **14**, 317–384 (with discussion).

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1**, 385–402 and 457–464, (with discussion).

Bernardo, J. M. (2007). Objective Bayesian point and region estimation in location-scale models. *Sort* **31**, 3–44, (with discussion).

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37**, 905–938.

Bernardo, J. M. (2011). Integrated objective Bayesian estimation and hypothesis testing. *Bayesian Statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press (to appear).

Berger, J. O., Bernardo, J. M. and Sun, D. (2011). Reference priors for discrete parameters. *J. Amer. Statist. Assoc.* (under revision).