On alternative perspectives and solutions on Bayesian tests

Christian P. Robert
Université Paris-Dauphine, Paris & University of Warwick, Coventry

bayesianstatistics@gmail.com

Outline

Significance tests: one new parameter

Jeffreys-Lindley paradox

Deviance (information criterion)

Aitkin's integrated likelihood

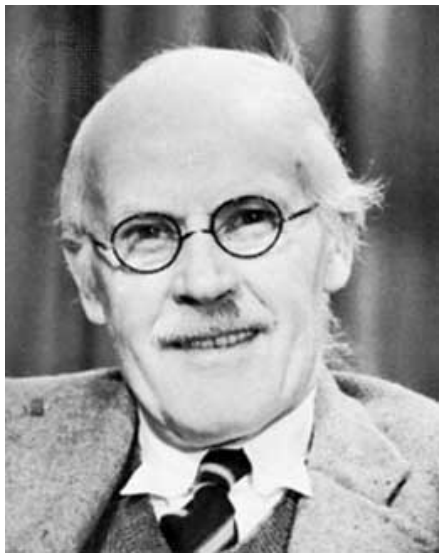
Johnson's uniformly most powerful Bayesian tests

Posterior predictive checking

“Significance tests: one new parameter”

Significance tests: one new parameter

- Bayesian tests
- Bayes factors
- Improper priors for tests
- Conclusion



Jeffreys-Lindley paradox

Deviance (information criterion)

Aitkin’s integrated likelihood

Johnson’s uniformly most powerful Bayesian tests

Posterior predictive checking

Fundamental setting

Is the new parameter supported by the observations or is any variation expressible by it better interpreted as random? Thus we must set two hypotheses for comparison, the more complicated having the smaller initial probability (Jeffreys, ToP, V, §5.0)

...compare a specially suggested value of a new parameter, often 0 [q], with the aggregate of other possible values [q']. We shall call q the null hypothesis and q' the alternative hypothesis [and] we must take

$$P(q|H) = P(q'|H) = 1/2\,.$$

Construction of Bayes tests

Definition (Test)

Given an hypothesis $H_0 : \theta \in \Theta_0$ on the parameter $\theta \in \Theta_0$ of a statistical model, a **test** is a statistical procedure that takes its values in $\{0, 1\}$.

Type–one and type–two errors

Associated with the risk

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[\mathsf{aL}(\theta, \delta(x))] \\ &= \begin{cases} P_\theta(\delta(x) = 0) & \text{if } \theta \in \Theta_0, \\ P_\theta(\delta(x) = 1) & \text{otherwise,} \end{cases} \end{aligned}$$

Theorem (Bayes test)

The Bayes estimator associated with π and with the $0-1$ loss is

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } P(\theta \in \Theta_0 | x) > P(\theta \notin \Theta_0 | x), \\ 0 & \text{otherwise,} \end{cases}$$

Type–one and type–two errors

Associated with the risk

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[\mathrm{aL}(\theta, \delta(x))] \\ &= \begin{cases} P_\theta(\delta(x) = 0) & \text{if } \theta \in \Theta_0, \\ P_\theta(\delta(x) = 1) & \text{otherwise,} \end{cases} \end{aligned}$$

Theorem (Bayes test)

The Bayes estimator associated with π and with the $0-1$ loss is

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } P(\theta \in \Theta_0 | x) > P(\theta \notin \Theta_0 | x), \\ 0 & \text{otherwise,} \end{cases}$$

Jeffreys' example (§5.0)

Testing whether the mean α of a normal observation is zero:

$$P(q|aH) \propto \exp\left(-\frac{a^2}{2s^2}\right)$$

$$P(q'\mathrm{d}\alpha|aH) \propto \exp\left(-\frac{(a-\alpha)^2}{2s^2}\right) f(\alpha)\mathrm{d}\alpha$$

$$P(q'|aH) \propto \int \exp\left(-\frac{(a-\alpha)^2}{2s^2}\right) f(\alpha)\mathrm{d}\alpha$$

A (small) point of contention

Jeffreys asserts

> *Suppose that there is one old parameter α; the new parameter is β and is 0 on q. In q' we could replace α by α', any function of α and β: but to make it explicit that q' reduces to q when $\beta = 0$ we shall require that $\alpha' = \alpha$ when $\beta = 0$ (V, §5.0).*

This amounts to assume identical parameters in both models, a controversial principle for model choice or at the very best to make α and β dependent a priori, a choice contradicted by the next paragraph in **ToP**

A (small) point of contention

Jeffreys asserts

> *Suppose that there is one old parameter α; the new parameter is β and is 0 on q. In q' we could replace α by α', any function of α and β: but to make it explicit that q' reduces to q when $\beta = 0$ we shall require that $\alpha' = \alpha$ when $\beta = 0$ (V, §5.0).*

This amounts to assume identical parameters in both models, a controversial principle for model choice or at the very best to make α and β dependent a priori, a choice contradicted by the next paragraph in **ToP**

Orthogonal parameters

If

$$I(\alpha, \beta) = \begin{bmatrix} g_{\alpha\alpha} & 0 \\ 0 & g_{\beta\beta} \end{bmatrix},$$

α and β *orthogonal*, but not [a posteriori] *independent*, contrary to **ToP** assertions

> *...the result will be nearly independent on previous information on old parameters (V, §5.01).*

and

$$K = \frac{1}{f(b,a)} \sqrt{\frac{n g_{\beta\beta}}{2\pi}} \exp\left(-\frac{1}{2} n g_{\beta\beta} b^2\right)$$

> *[where] $h(\alpha)$ is irrelevant (V, §5.01)*

Orthogonal parameters

If

$$I(\alpha, \beta) = \begin{bmatrix} g_{\alpha\alpha} & 0 \\ 0 & g_{\beta\beta} \end{bmatrix},$$

α and β *orthogonal*, but not [a posteriori] *independent*, contrary to **ToP** assertions

> *...the result will be nearly independent on previous information on old parameters (V, §5.01).*

and

$$K = \frac{1}{f(b, a)} \sqrt{\frac{n g_{\beta\beta}}{2\pi}} \exp\left(-\frac{1}{2} n g_{\beta\beta} b^2\right)$$

> *[where] $h(\alpha)$ is irrelevant (V, §5.01)*

Acknowledgement in **ToP**

In practice it is rather unusual for a set of parameters to arise in such a way that each can be treated as irrelevant to the presence of any other. More usual cases are (...) where some parameters are so closely associated that one could hardly occur without the others (V, §5.04).

Generalisation

Theorem (Optimal Bayes decision)

Under the $0-1$ *loss function*

$$L(\theta, d) = \begin{cases} 0 & \text{if } d = \mathbb{I}_{\Theta_0}(\theta) \\ a_0 & \text{if } d = 1 \text{ and } \theta \notin \Theta_0 \\ a_1 & \text{if } d = 0 \text{ and } \theta \in \Theta_0 \end{cases}$$

the Bayes procedure is

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } \Pr^\pi(\theta \in \Theta_0 | x) \geqslant a_0/(a_0 + a_1) \\ 0 & \text{otherwise} \end{cases}$$

Generalisation

Theorem (Optimal Bayes decision)

Under the $0-1$ *loss function*

$$L(\theta, d) = \begin{cases} 0 & \text{if } d = \mathbb{I}_{\Theta_0}(\theta) \\ a_0 & \text{if } d = 1 \text{ and } \theta \notin \Theta_0 \\ a_1 & \text{if } d = 0 \text{ and } \theta \in \Theta_0 \end{cases}$$

the Bayes procedure is

$$\delta^{\pi}(x) = \begin{cases} 1 & \text{if } \Pr^{\pi}(\theta \in \Theta_0 | x) \geqslant a_0/(a_0 + a_1) \\ 0 & \text{otherwise} \end{cases}$$

Bound comparison

Determination of a_0/a_1 depends on consequences of "wrong decision" under both circumstances

Often difficult to assess in practice and replacement with "golden" default bounds like .05, biased towards H_0

Bound comparison

Determination of a_0/a_1 depends on consequences of "wrong decision" under both circumstances
Often difficult to assess in practice and replacement with "golden" default bounds like .05, biased towards H_0

A function of posterior probabilities

Definition (Bayes factors)

For hypotheses $H_0 : \theta \in \Theta_0$ vs. $H_a : \theta \notin \Theta_0$

$$\mathfrak{B}_{01} = \frac{\pi(\Theta_0|x)}{\pi(\Theta_0^c|x)} \Big/ \frac{\pi(\Theta_0)}{\pi(\Theta_0^c)} = \frac{\displaystyle\int_{\Theta_0} f(x|\theta)\pi_0(\theta)\mathrm{d}\theta}{\displaystyle\int_{\Theta_0^c} f(x|\theta)\pi_1(\theta)\mathrm{d}\theta}$$

[Good, 1958 & **ToP**, V, §5.01]

Equivalent to Bayes rule: acceptance if

$$\mathfrak{B}_{01} > \{(1-\pi(\Theta_0))/a_1\}/\{\pi(\Theta_0)/a_0\}$$

A major modification

When the null hypothesis is supported by a set of measure 0 against Lebesgue measure, $\pi(\Theta_0) = 0$ for an absolutely continuous prior distribution

[End of the story?!]

> *Suppose we are considering whether a location parameter α is 0. The estimation prior probability for it is uniform and we should have to take $f(\alpha) = 0$ and $K[= \mathfrak{B}_{10}]$ would always be infinite (V, §5.02)*

A major modification

When the null hypothesis is supported by a set of measure 0 against Lebesgue measure, $\pi(\Theta_0) = 0$ for an absolutely continuous prior distribution

[End of the story?!]

> *Suppose we are considering whether a location parameter* α *is* 0. *The estimation prior probability for it is uniform and we should have to take* $f(\alpha) = 0$ *and* $K[= \mathfrak{B}_{10}]$ *would always be infinite (V,* §*5.02)*

Point null refurbishment

Requirement

Defined prior distributions under both assumptions,

$$\pi_0(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_0}(\theta), \quad \pi_1(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_1}(\theta),$$

(under the standard dominating measures on Θ_0 and Θ_1)

Using the prior probabilities $\pi(\Theta_0) = \rho_0$ and $\pi(\Theta_1) = \rho_1$,

$$\pi(\theta) = \rho_0\pi_0(\theta) + \rho_1\pi_1(\theta).$$

Note If $\Theta_0 = \{\theta_0\}$, π_0 is the Dirac mass in θ_0

Point null refurbishment

Requirement

Defined prior distributions under both assumptions,

$$\pi_0(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_0}(\theta), \quad \pi_1(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_1}(\theta),$$

(under the standard dominating measures on Θ_0 and Θ_1)

Using the prior probabilities $\pi(\Theta_0) = \rho_0$ and $\pi(\Theta_1) = \rho_1$,

$$\pi(\theta) = \rho_0\pi_0(\theta) + \rho_1\pi_1(\theta).$$

Note If $\Theta_0 = \{\theta_0\}$, π_0 is the Dirac mass in θ_0

Point null hypotheses

Particular case $H_0 : \theta = \theta_0$
Take $\rho_0 = \Pr^{\pi}(\theta = \theta_0)$ and g_1 prior density under H_a.
Posterior probability of H_0

$$\pi(\Theta_0|x) = \frac{f(x|\theta_0)\rho_0}{\int f(x|\theta)\pi(\theta)\,d\theta} = \frac{f(x|\theta_0)\rho_0}{f(x|\theta_0)\rho_0 + (1-\rho_0)m_1(x)}$$

and marginal under H_a

$$m_1(x) = \int_{\Theta_1} f(x|\theta)g_1(\theta)\,d\theta.$$

Point null hypotheses

Particular case $H_0 : \ \theta = \theta_0$
Take $\rho_0 = \Pr^{\pi}(\theta = \theta_0)$ and g_1 prior density under H_a.
Posterior probability of H_0

$$\pi(\Theta_0|x) = \frac{f(x|\theta_0)\rho_0}{\int f(x|\theta)\pi(\theta)\, d\theta} = \frac{f(x|\theta_0)\rho_0}{f(x|\theta_0)\rho_0 + (1-\rho_0)m_1(x)}$$

and marginal under H_a

$$m_1(x) = \int_{\Theta_1} f(x|\theta)g_1(\theta)\, d\theta.$$

Point null hypotheses (cont'd)

Dual representation

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}.$$

and

$$\mathfrak{B}^{\pi}_{01}(x) = \frac{f(x|\theta_0)\rho_0}{m_1(x)(1-\rho_0)} \Big/ \frac{\rho_0}{1-\rho_0} = \frac{f(x|\theta_0)}{m_1(x)}$$

Connection

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{1}{\mathfrak{B}^{\pi}_{01}(x)}\right]^{-1}.$$

Point null hypotheses (cont'd)

Dual representation

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}.$$

and

$$\mathfrak{B}^{\pi}_{01}(x) = \frac{f(x|\theta_0)\rho_0}{m_1(x)(1-\rho_0)} \Big/ \frac{\rho_0}{1-\rho_0} = \frac{f(x|\theta_0)}{m_1(x)}$$

Connection

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{1}{\mathfrak{B}^{\pi}_{01}(x)}\right]^{-1}.$$

A further difficulty

Improper priors are not allowed here

If

$$\int_{\Theta_1} \pi_1(\mathrm{d}\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(\mathrm{d}\theta_2) = \infty$$

then π_1 or π_2 cannot be coherently normalised **while** the normalisation matters in the Bayes factor

▸ remember Bayes factor?

A further difficulty

Improper priors are not allowed here

If

$$\int_{\Theta_1} \pi_1(\mathrm{d}\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(\mathrm{d}\theta_2) = \infty$$

then π_1 or π_2 cannot be coherently normalised **while** the normalisation matters in the Bayes factor

▸ remember Bayes factor?

ToP unaware of the problem?

A. Not entirely, as improper priors keep being used on nuisance parameters

Example of testing for a zero normal mean:

> *If σ is the standard error and λ the true value, λ is 0 on q. We want a suitable form for its prior on q'. (...) Then we should take*
>
> $$P(q\,d\sigma|H) \propto d\sigma/\sigma$$
> $$P(q'\,d\sigma\,d\lambda|H) \propto f\left(\frac{\lambda}{\sigma}\right) d\sigma/\sigma\,d\lambda/\lambda$$
>
> *where f [is a true density] (V, §5.2).*

Fallacy of the "same" σ!

ToP unaware of the problem?

A. Not entirely, as improper priors keep being used on nuisance parameters

Example of testing for a zero normal mean:

> *If σ is the standard error and λ the true value, λ is 0 on q. We want a suitable form for its prior on q'. (...) Then we should take*
>
> $$P(q\,d\sigma|H) \propto d\sigma/\sigma$$
> $$P(q'\,d\sigma\,d\lambda|H) \propto f\left(\frac{\lambda}{\sigma}\right) d\sigma/\sigma\, d\lambda/\lambda$$
>
> *where f [is a true density] (V, §5.2).*

Fallacy of the "same" σ!

Not enought information

If $s' = 0$ *[!!!], then [for* $\sigma = |\bar{x}|/\tau$, $\lambda = \sigma v$*]*

$$P(q|\theta H) \propto \int_0^\infty \left(\frac{\tau}{|\bar{x}|}\right)^n \exp\left(-\frac{1}{2}n\tau^2\right)\frac{d\tau}{\tau},$$

$$P(q'|\theta H) \propto \int_0^\infty \frac{d\tau}{\tau}\int_{-\infty}^\infty \left(\frac{\tau}{|\bar{x}|}\right)^n f(v)\exp\left(-\frac{1}{2}n(v-\tau)^2\right).$$

If $n = 1$ *and* $f(v)$ *is any even [density],*

$$P(q'|\theta H) \propto \frac{1}{2}\frac{\sqrt{2\pi}}{|\bar{x}|} \quad \textit{and} \quad P(q|\theta H) \propto \frac{1}{2}\frac{\sqrt{2\pi}}{|\bar{x}|}$$

and therefore $K = 1$ *(V, §5.2).*

Strange constraints

If $n \geqslant 2$, the condition that $K = 0$ for $s' = 0$, $\bar{x} \neq 0$ is equivalent to

$$\int_0^\infty f(v) v^{n-1} dv = \infty \,.$$

The function satisfying this condition for [all] n is

$$f(v) = \frac{1}{\pi(1+v^2)}$$

This is the prior recommended by Jeffreys hereafter.
But, first, many other families of densities satisfy this constraint and a scale of 1 cannot be **universal**!
Second, $s' = 0$ is a zero probability event...

Strange constraints

If $n \geqslant 2$, the condition that $K = 0$ for $s' = 0$, $\bar{x} \neq 0$ is equivalent to

$$\int_0^\infty f(v) v^{n-1} dv = \infty \,.$$

The function satisfying this condition for [all] n is

$$f(v) = \frac{1}{\pi(1+v^2)}$$

This is the prior recommended by Jeffreys hereafter.
But, first, many other families of densities satisfy this constraint and a scale of 1 cannot be **universal**!
Second, $s' = 0$ is a zero probability event...

Strange constraints

If $n \geqslant 2$, the condition that $K = 0$ for $s' = 0$, $\bar{x} \neq 0$ is equivalent to

$$\int_0^\infty f(v) v^{n-1} dv = \infty \,.$$

The function satisfying this condition for [all] n is

$$f(v) = \frac{1}{\pi(1+v^2)}$$

This is the prior recommended by Jeffreys hereafter.
But, first, many other families of densities satisfy this constraint and a scale of 1 cannot be **universal**!
Second, $s' = 0$ is a zero probability event...

Comments

- **ToP** very imprecise about choice of priors in the setting of tests (despite existence of Susie's Jeffreys' conventional partly proper priors)
- **ToP** misses the difficulty of improper priors [coherent with earlier stance]
- but this problem still generates debates within the B community
- Some degree of goodness-of-fit testing but against fixed alternatives
- Persistence of the form

$$K \approx \sqrt{\frac{\pi n}{2}} \left(1 + \frac{t^2}{\nu}\right)^{-1/2\nu+1/2}$$

but ν not so clearly defined...

Jeffreys–Lindley paradox

Significance tests: one new parameter

Jeffreys-Lindley paradox
- Lindley's paradox
- dual versions of the paradox
- "Who should be afraid of the Lindley–Jeffreys paradox?"
- Bayesian resolutions

Deviance (information criterion)

Aitkin's integrated likelihood

Johnson's uniformly most powerful Bayesian tests

Lindley's paradox

In a normal mean testing problem,

$$\bar{x}_n \sim \mathcal{N}(\theta, \sigma^2/n)\,, \quad H_0 : \theta = \theta_0\,,$$

under Jeffreys prior, $\theta \sim \mathcal{N}(\theta_0, \sigma^2)$, the Bayes factor

$$\mathfrak{B}_{01}(t_n) = (1+n)^{1/2} \exp\left(-nt_n^2/2[1+n]\right)\,,$$

where $t_n = \sqrt{n}|\bar{x}_n - \theta_0|/\sigma$, satisfies

$$\mathfrak{B}_{01}(t_n) \overset{n\longrightarrow\infty}{\longrightarrow} \infty$$

[assuming a fixed t_n]

[Lindley, 1957]

Lindley's paradox

Often dubbed *Jeffreys–Lindley paradox...*

In terms of

$$t = \sqrt{n-1}\,\bar{x}/s', \qquad \nu = n-1$$

$$K \sim \sqrt{\frac{\pi\nu}{2}}\left(1+\frac{t^2}{\nu}\right)^{-1/2\nu+1/2}.$$

(...) The variation of K with t is much more important than the variation with ν (Jeffreys, V, §5.2).

Two versions of the paradox

"the weight of Lindley's paradoxical result (...) burdens proponents of the Bayesian practice".

[Lad, 2003]

- official version, opposing frequentist and Bayesian assessments

[Lindley, 1957]

- intra-Bayesian version, blaming vague and improper priors for the Bayes factor misbehaviour:
  if $\pi_1(\cdot|\sigma)$ depends on a scale parameter σ, it is often the case that

$$\mathfrak{B}_{01}(x) \overset{\sigma\longrightarrow\infty}{\longrightarrow} +\infty$$

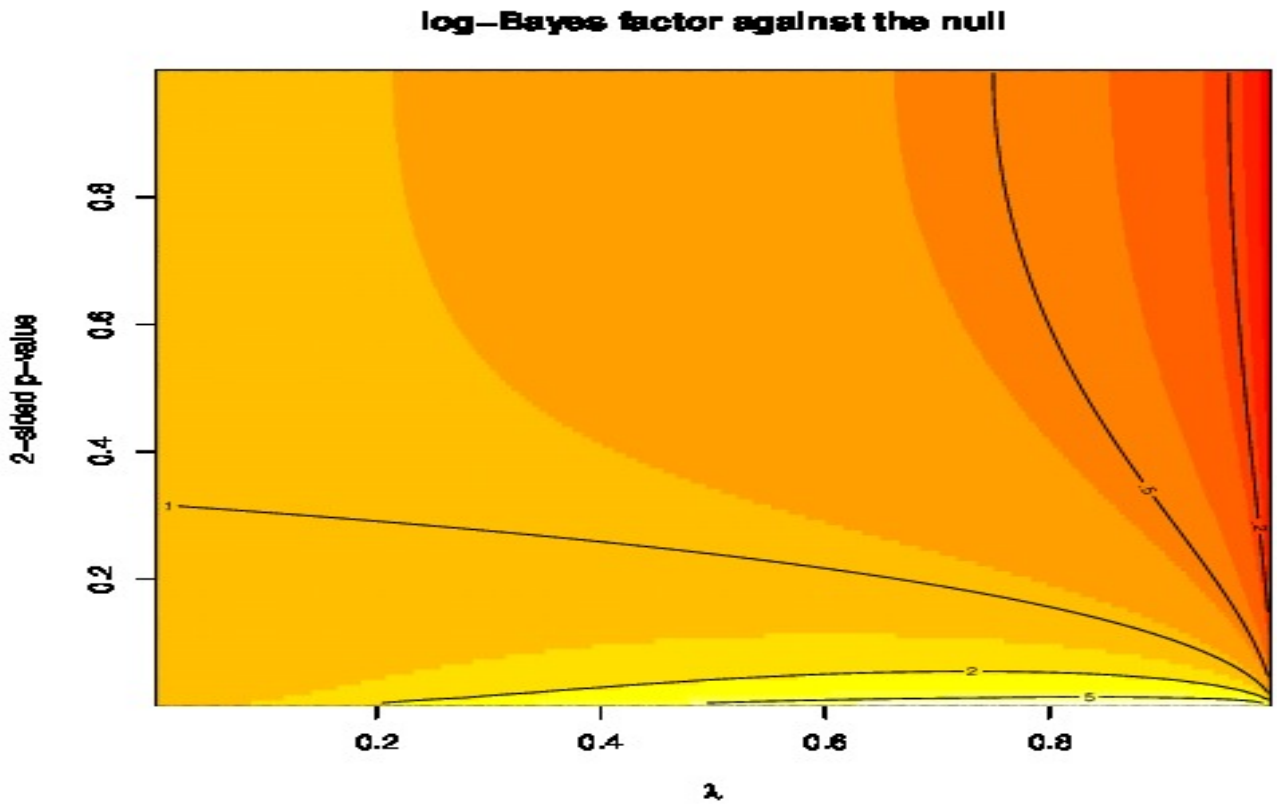
  for a given x, meaning H_0 is always accepted

[Robert, 1992, 2013]

where does it matter?

In the normal case, $Z \sim \mathcal{N}(\theta, 1)$, $\theta \sim \mathcal{N}(0, \alpha^2)$, Bayes factor

$$\mathfrak{B}_{10}(z) = \frac{e^{z^2\alpha^2/(1+\alpha^2)}}{\sqrt{1+\alpha^2}} = \sqrt{1-\lambda}\,\exp\{\lambda z^2/2\}$$



Evacuation of the first version

Two paradigms [(𝔟) versus (𝔣)]

- one (𝔟) operates on the parameter space Θ, while the other (𝔣) is produced from the sample space
- one (𝔣) relies solely on the point-null hypothesis H_0 and the corresponding sampling distribution, while the other (𝔟) opposes H_0 to a (predictive) marginal version of H_1
- one (𝔣) could reject *"a hypothesis that may be true (...) because it has not predicted observable results that have not occurred" (Jeffreys,* **ToP***, VII, §7.2)* while the other (𝔟) conditions upon the observed value x_{obs}
- one (𝔣) cannot agree with the likelihood principle, while the other (𝔟) is almost uniformly in agreement with it
- one (𝔣) resorts to an arbitrary fixed bound α on the p-value, while the other (𝔟) refers to the (default) boundary probability of $1/2$

More arguments on the first version

- observing a constant t_n as n increases is of limited interest: under H_0 t_n has limiting $\mathcal{N}(0,1)$ distribution, while, under H_1 t_n a.s. converges to ∞
- behaviour that remains entirely compatible with the consistency of the Bayes factor, which a.s. converges either to 0 or ∞, depending on which hypothesis is true.

Consequent subsequent literature (e.g., Berger & Sellke, 1987; Bayarri & Berger, 2004) has since then shown how divergent those two approaches could be (to the point of being asymptotically incompatible).

Nothing's wrong with the second version

- n, prior's scale factor: prior variance n times larger than the observation variance and *when n goes to ∞, Bayes factor goes to ∞ no matter what the observation is*
- n becomes what Lindley (1957) calls *"a measure of lack of conviction about the null hypothesis"*
- when prior diffuseness under H_1 increases, only relevant information becomes that θ could be equal to θ_0, and this overwhelms any evidence to the contrary contained in the data
- mass of the prior distribution in the vicinity of any fixed neighbourhood of the null hypothesis vanishes to zero under H_1

© **deep coherence in the outcome: being indecisive about the alternative hypothesis means we should not chose it**

Nothing's wrong with the second version

- n, prior's scale factor: prior variance n times larger than the observation variance and *when n goes to ∞, Bayes factor goes to ∞ no matter what the observation is*
- n becomes what Lindley (1957) calls *"a measure of lack of conviction about the null hypothesis"*
- when prior diffuseness under H_1 increases, only relevant information becomes that θ could be equal to θ_0, and this overwhelms any evidence to the contrary contained in the data
- mass of the prior distribution in the vicinity of any fixed neighbourhood of the null hypothesis vanishes to zero under H_1

© **deep coherence in the outcome: being indecisive about the alternative hypothesis means we should not chose it**

"Who should be afraid of the Lindley–Jeffreys paradox?"

Recent publication by A. Spanos with above title:

- the paradox demonstrates against Bayesian and likelihood resolutions of the problem for failing to account for the large sample size.
- the failure of all three main paradigms ("fallacy of rejection" for (𝔣) versus "fallacy of acceptance" for (𝔟)) leads to advocate Mayo's and Spanos' (2004) "postdata severity evaluation"

[Spanos, 2013]

"Who should be afraid of the Lindley–Jeffreys paradox?"

Recent publication by A. Spanos with above title:

> *"the postdata severity evaluation (...) addresses the key problem with Fisherian p-values in the sense that the severity evaluation provides the "magnitude" of the warranted discrepancy from the null by taking into account the generic capacity of the test (that includes n) in question as it relates to the observed data"(p.88)*

[Spanos, 2013]

what is severity?

"An hypothesis H passes a severe test if the data agrees with H and if it is highly probable that data not produced under H agrees less with H"

- departure from the null, rewritten as $\theta_1 = \theta_0 + \gamma$,
- "provide the 'magnitude' of the warranted discrepancy from the null", i.e. decide about how close (in distance) to the null we can get and still be able to discriminate the null from the alternative hypotheses "with very high probability"
- requires to set the "severity threshold",

$$\mathbb{P}_{\theta_1}\{d(\mathbf{X}) \leqslant d(x_0)\}$$

- once γ found, whether it is far enough from the null is a matter of informed opinion: whether it is "substantially significant (...) pertains to the substantive subject matter"

...should we be afraid?

A. Not! In Spanos (2013)

- the purpose of a test and the nature of evidence are never spelled out
- the rejection of decisional aspects clashes with the later call to the magnitude of the severity
- does not quantify how to select significance thresholds γ against sample size n
- contains irrelevant attacks on the likelihood principle and dependence on Euclidean distance

[Robert, 2013]

On some resolutions of the second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc, which lacks complete proper Bayesian justification

  [Berger & Pericchi, 2001]

- use of *identical* improper priors on nuisance parameters,
- use of the posterior predictive distribution,
- matching priors,
- use of score functions extending the log score function
- non-local priors correcting default priors

On some resolutions of the second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters, a notion already entertained by Jeffreys

[Berger et al., 1998; Marin & Robert, 2013]

- use of the posterior predictive distribution,
- matching priors,
- use of score functions extending the log score function
- non-local priors correcting default priors

On some resolutions of the second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- *Péché de jeunesse:* equating the values of the prior densities at the point-null value θ_0,

$$\rho_0 = (1 - \rho_0)\pi_1(\theta_0)$$

[Robert, 1993]

- use of the posterior predictive distribution,
- matching priors,
- use of score functions extending the log score function
- non-local priors correcting default priors

On some resolutions of the second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- use of the posterior predictive distribution, which uses the data twice
- matching priors,
- use of score functions extending the log score function
- non-local priors correcting default priors

On some resolutions of the second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- use of the posterior predictive distribution,
- matching priors, whose sole purpose is to bring frequentist and Bayesian coverages as close as possible

  [Datta & Mukerjee, 2004]
- use of score functions extending the log score function
- non-local priors correcting default priors

On some resolutions of the second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- use of the posterior predictive distribution,
- matching priors,
- use of score functions extending the log score function

$$\log \mathfrak{B}_{12}(x) = \log m_1(x) - \log m_2(x) = S_0(x, m_1) - S_0(x, m_2)\,,$$

that are independent of the normalising constant

[Dawid et al., 2013]

- non-local priors correcting default priors

On some resolutions of the second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- use of the posterior predictive distribution,
- matching priors,
- use of score functions extending the log score function
- non-local priors correcting default priors towards more balanced error rates

[Johnson & Rossell, 2010; Consonni et al., 2013]

Deviance (information criterion)

Significance tests: one new parameter

Jeffreys-Lindley paradox

Deviance (information criterion)

Aitkin's integrated likelihood

Johnson's uniformly most powerful Bayesian tests

Posterior predictive checking

DIC as in Dayesian?

Deviance defined by

$$D(\theta) = -2\log(p(\mathbf{y}|\theta))\,,$$

Effective number of parameters computed as

$$p_D = \bar{D} - D(\bar{\theta})\,,$$

with $\bar{D}$ posterior expectation of D and $\bar{\theta}$ estimate of θ
Deviance information criterion (DIC) defined by

$$\begin{aligned}\mathrm{DIC} &= p_D + \bar{D} \\ &= D(\bar{\theta}) + 2p_D\end{aligned}$$

Models with smaller DIC better supported by the data

[Spiegelhalter et al., 2002]

"thou shalt not use the data twice"

The data is used twice in the DIC method:

1. y used **once** to produce the posterior $\pi(\theta|y)$, and the associated estimate, $\tilde{\theta}(y)$
2. y used **a second time** to compute the posterior expectation of the *observed* likelihood $p(y|\theta)$,

$$\int \log p(y|\theta)\pi(\mathrm{d}\theta|y) \propto \int \log p(y|\theta)p(y|\theta)\pi(\mathrm{d}\theta)\,,$$

DIC for missing data models

Framework of missing data models

$$f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \boldsymbol{z}|\theta) d\boldsymbol{z}\,,$$

with observed data $\mathbf{y} = (y_1, \ldots, y_n)$ and corresponding *missing data* by $\boldsymbol{z} = (z_1, \ldots, z_n)$

How do we define DIC in such settings?

DIC for missing data models

Framework of missing data models

$$f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \boldsymbol{z}|\theta) d\boldsymbol{z}\,,$$

with observed data $\mathbf{y} = (y_1, \ldots, y_n)$ and corresponding *missing data* by $\boldsymbol{z} = (z_1, \ldots, z_n)$

How do we define DIC in such settings?

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs

$$\text{DIC}_1 = -4\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\mathbb{E}_\theta\left[\theta|\mathbf{y}\right])$$

   often a poor choice in case of unidentifiability
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs

$$\text{DIC}_2 = -4\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\widehat{\theta}(\mathbf{y}))\,.$$

   which uses posterior mode instead
2. complete DICs based on $f(\mathbf{y},\mathbf{z}|\theta)$
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z},\theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs

$$\text{DIC}_3 = -4\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log \widehat{f}(\mathbf{y})\,,$$

   which instead relies on the MCMC density estimate
2. complete DICs based on $f(\mathbf{y},\mathbf{z}|\theta)$
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z},\theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$

$$\begin{aligned} \text{DIC}_4 &= \mathbb{E}_{\mathbf{Z}}\left[\text{DIC}(\mathbf{y}, \mathbf{Z})|\mathbf{y}\right] \\ &= -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}\right] + 2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\mathbb{E}_{\theta}[\theta|\mathbf{y}, \mathbf{Z}])|\mathbf{y}\right] \end{aligned}$$

3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$

$$\text{DIC}_5 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}, \widehat{\mathbf{z}}(\mathbf{y})|\widehat{\theta}(\mathbf{y}))\,,$$

   using $\mathbf{Z}$ as an additional parameter
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$

   $$\text{DIC}_6 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}\right] + 2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\widehat{\theta}(\mathbf{y}))|\mathbf{y}, \widehat{\theta}(\mathbf{y})\right].$$

   in analogy with EM, $\widehat{\theta}$ being an EM fixed point
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

$$\text{DIC}_7 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}|\mathbf{Z}, \theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\widehat{\mathbf{z}}(\mathbf{y}), \widehat{\theta}(\mathbf{y}))\,,$$

using MAP estimates

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

$$\text{DIC}_8 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}|\mathbf{Z}, \theta)|\mathbf{y}\right] + 2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y}|\mathbf{Z}, \widehat{\theta}(\mathbf{y}, \mathbf{Z}))|\mathbf{y}\right] ,$$

conditioning first on $\mathbf{Z}$ and then integrating over $\mathbf{Z}$ conditional on $\mathbf{y}$

[Celeux et al., BA, 2006]

Galactic DICs

Example of the galaxy mixture dataset

K	DIC_2 (p_{D2})	DIC_3 (p_{D3})	DIC_4 (p_{D4})	DIC_5 (p_{D5})	DIC_6 (p_{D6})	DIC_7 (p_{D7})	DIC_8 (p_{D8})
2	453	451	502	705	501	417	410
	(5.56)	(3.66)	(5.50)	(207.88)	(4.48)	(11.07)	(4.09)
3	440	436	461	622	471	378	372
	(9.23)	(4.94)	(6.40)	(167.28)	(15.80)	(13.59)	(7.43)
4	446	439	473	649	482	388	382
	(11.58)	(5.41)	(7.52)	(183.48)	(16.51)	(17.47)	(11.37)
5	447	442	485	658	511	395	390
	(10.80)	(5.48)	(7.58)	(180.73)	(33.29)	(20.00)	(15.15)
6	449	444	494	676	532	407	398
	(11.26)	(5.49)	(8.49)	(191.10)	(46.83)	(28.23)	(19.34)
7	460	446	508	700	571	425	409
	(19.26)	(5.83)	(8.93)	(200.35)	(71.26)	(40.51)	(24.57)

questions

- what is the behaviour of DIC under model mispecification?
- is there an absolute scale to the DIC values, i.e. when is a difference in DICs significant?
- how can DIC handle small n's versus p's?
- should p_D be defined as $\text{var}(D|\mathbf{y})/2$ [Gelman's suggestion]?
- is WAIC (Gelman and Vehtari, 2013) making a difference for being based on expected posterior predictive?

In an era of complex models, is DIC applicable?

[Robert, 2013]

questions

- what is the behaviour of DIC under model mispecification?
- is there an absolute scale to the DIC values, i.e. when is a difference in DICs significant?
- how can DIC handle small n's versus p's?
- should p_D be defined as $\text{var}(D|\mathbf{y})/2$ [Gelman's suggestion]?
- is WAIC (Gelman and Vehtari, 2013) making a difference for being based on expected posterior predictive?

In an era of complex models, is DIC applicable?

[Robert, 2013]

Aitkin's integrated likelihood

Significance tests: one new parameter

Jeffreys-Lindley paradox

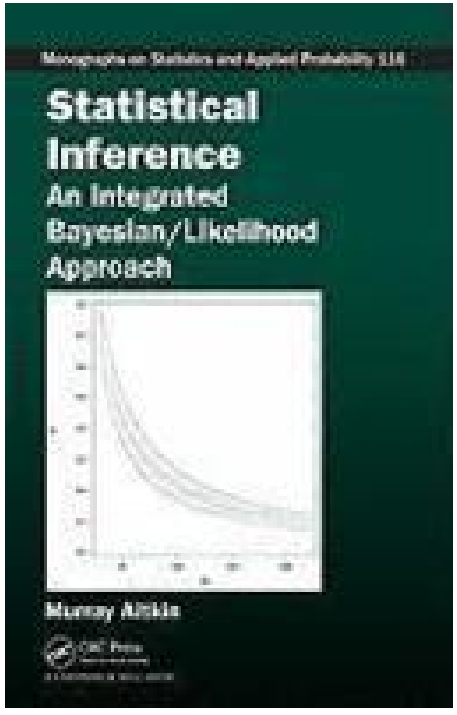
Deviance (information criterion)

Aitkin's integrated likelihood

- Integrated likelihood
- Criticisms
- A Bayesian version?

Johnson's uniformly most powerful Bayesian tests

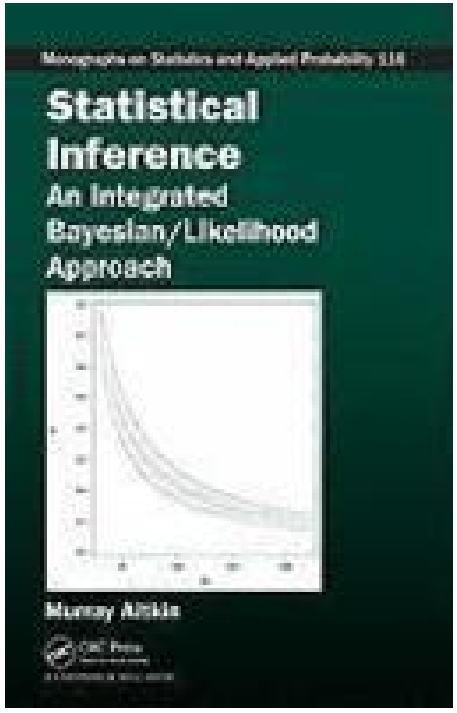
Posterior predictive checking



Integrated likelihood

Statistical Inference: An Integrated Bayesian/Likelihood Approach was published by Murray Aitkin in 2009

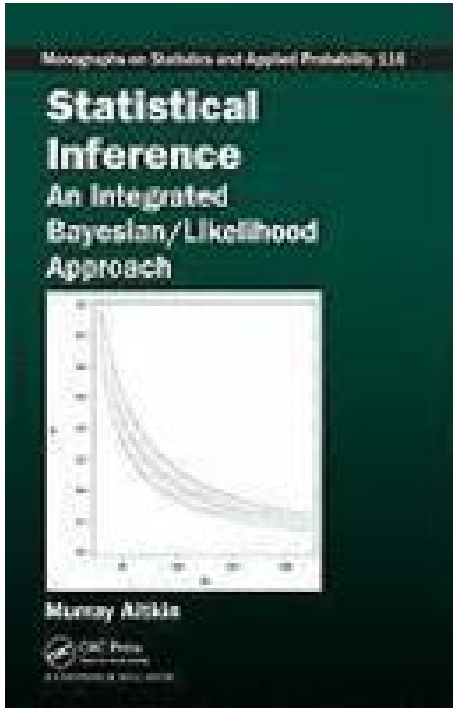
Theme: comparisons of posterior distributions of likelihood functions under competing models or via the posterior distribution of likelihood ratios corresponding to those models...



Integrated likelihood

Statistical Inference: An Integrated Bayesian/Likelihood Approach was published by Murray Aitkin in 2009

Theme: comparisons of posterior distributions of likelihood functions under competing models or via the posterior distribution of likelihood ratios corresponding to those models...



Posterior likelihood

"This quite small change to standard Bayesian analysis allows a very general approach to a wide range of apparently different inference problems; a particular advantage of the approach is that it can use the same noninformative priors." Statistical Inference, p.xiii

Central tool: "posterior cdf" of the likelihood,

$$F(z) = \Pr^{\pi}(L(\theta, x) > z|x)\,.$$

Arguments:

- general approach that resolves difficulties with the Bayesian processing of point null hypotheses
- includes use of generic noninformative and improper priors
- handles the "vexed question of model fit"

Posterior likelihood

> *"This quite small change to standard Bayesian analysis allows a very general approach to a wide range of apparently different inference problems; a particular advantage of the approach is that it can use the same noninformative priors." Statistical Inference, p.xiii*

Central tool: "posterior cdf" of the likelihood,

$$F(z) = \Pr^{\pi}(L(\theta, x) > z|x)\,.$$

Arguments:

- general approach that resolves difficulties with the Bayesian processing of point null hypotheses
- includes use of generic noninformative and improper priors
- handles the "vexed question of model fit"

Using the data twice [again!]

"A persistent criticism of the posterior likelihood approach (...) has been based on the claim that these approaches are 'using the data twice,' or are 'violating temporal coherence'." Statistical Inference, p.48

- "posterior expectation" of the likelihood as ratio of marginal of twice-replicated data over marginal of original data,

$$\mathbb{E}[L(\theta, x)|x] = \int L(\theta, x)\pi(\theta|x)\,\mathrm{d}\theta = \frac{m(x, x)}{m(x)}\,,$$

[Aitkin, 1991]

- the likelihood function does not exist a priori
- requires a *joint* distribution across models to be compared
- connection with pseudo-priors (Carlin & Chib, 1995) who defined prior distributions on the parameters that do not exist
- fails to include improper priors since (θ, x) has no joint distribution

Using the data twice [again!]

"A persistent criticism of the posterior likelihood approach (...) has been based on the claim that these approaches are 'using the data twice,' or are 'violating temporal coherence'." Statistical Inference, p.48

- "posterior expectation" of the likelihood as ratio of marginal of twice-replicated data over marginal of original data,

$$\mathbb{E}[L(\theta, x)|x] = \int L(\theta, x)\pi(\theta|x)\,\mathrm{d}\theta = \frac{m(x, x)}{m(x)}\,,$$

[Aitkin, 1991]

- the likelihood function does not exist a priori
- requires a *joint* distribution across models to be compared
- connection with pseudo-priors (Carlin & Chib, 1995) who defined prior distributions on the parameters that do not exist
- fails to include improper priors since (θ, x) has no joint distribution

Posterior probability on posterior probabilities

> *"The p-value is equal to the posterior probability that the likelihood ratio, for null hypothesis to alternative, is greater than 1 (...) The posterior probability is p that the posterior probability of H_0 is greater than 0.5."*
> *Statistical Inference, pp.42–43*

© A posterior probability being a *number*, how can its posterior probability be defined?

While

$$m(x) = \int L(\theta, x)\pi(\theta)\,\mathrm{d}\theta = \mathbb{E}^{\pi}[L(\theta, x)]$$

is well-defined, it does not mean the whole distribution of $L(\theta, x)$ makes sense!

Posterior probability on posterior probabilities

> *"The p-value is equal to the posterior probability that the likelihood ratio, for null hypothesis to alternative, is greater than 1 (...) The posterior probability is p that the posterior probability of H_0 is greater than 0.5."*
> *Statistical Inference, pp.42–43*

© A posterior probability being a *number*, how can its posterior probability be defined?

While

$$m(x) = \int L(\theta, x)\pi(\theta)\,\mathrm{d}\theta = \mathbb{E}^{\pi}[L(\theta, x)]$$

is well-defined, it does not mean the whole distribution of $L(\theta, x)$ makes sense!

Drifting apart

fundamental theoretical argument: integrated likelihood leads to *parallel and separate* simulations from the posteriors under each model, considering distribution of

$$L_i(\theta_i|x) \Big/ L_k(\theta_k|x),$$

when θ_i's and θ_k's drawn from respective posteriors

[see also Scott, 2002; Congdon, 2006]

Drifting apart

fundamental theoretical argument: integrated likelihood leads to *parallel and separate* simulations from the posteriors under each model, considering distribution of

$$L_i(\theta_i|x) \Big/ L_k(\theta_k|x),$$

when θ_i's and θ_k's drawn from respective posteriors

[see also Scott, 2002; Congdon, 2006]

MCMC simulations run for each model separately and resulting MCMC samples gathered together to produce posterior distribution of

$$\rho_i L(\theta_i|x) \Big/ \sum_k \rho_k L(\theta_k|x)\,,$$

which do not correspond to genuine Bayesian solutions

[Robert and Marin, 2008]

Drifting apart

fundamental theoretical argument: integrated likelihood leads to *parallel and separate* simulations from the posteriors under each model, considering distribution of

$$L_i(\theta_i|x) \Big/ L_k(\theta_k|x),$$

when θ_i's and θ_k's drawn from respective posteriors

[see also Scott, 2002; Congdon, 2006]

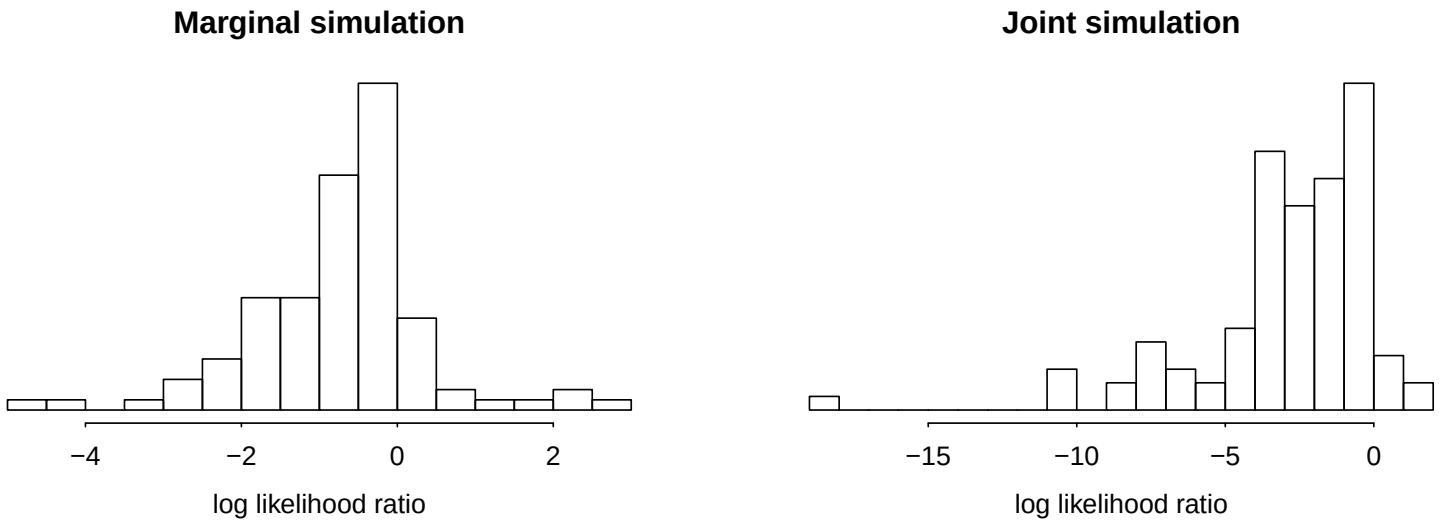
© the product of the posteriors $\pi_1(\theta_1|y^n)\pi_2(\theta_2|y^n)$ is not the posterior of the product $\pi(\theta_1, \theta_2|y^n)$, as in

$$p_1 m_1(x)\pi_1(\theta_1|x)\pi_2(\theta_2) + p_2 m_2(x)\pi_2(\theta_2|x)\pi_1(\theta_1).$$

[Carlin & Chib, 1995]

An illustration

Comparison of the distribution of the likelihood ratio under (a) true joint posterior and (b) product of posteriors, when assessing fit of a Poisson against binomial model with $m = 5$ trials, for the observation $x = 3$



Appropriate loss function

Estimation loss for model index j, the values of the parameters under both models and observation x:

$$L(\delta, (j, \theta_j, \theta_{-j})) = \mathbb{I}_{\delta=1}\mathbb{I}_{f_2(x|\theta_2)>f_1(x|\theta_1)} + \mathbb{I}_{\delta=2}\mathbf{I}_{f_2(x|\theta_2)<f_1(x|\theta_1)}$$

($\delta = j$ means model j is chosen, and $f_j(.|\theta_j)$ denotes likelihood under model j)

Under this loss, Bayes (optimal) solution

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } \Pr^\pi(f_2(x|\theta_2) < f_1(x|\theta_1)|x) > \frac{1}{2} \\ 2 & \text{otherwise,} \end{cases}$$

depends on *joint* posterior distribution on (θ_1, θ_2), thus differs from Aitkin's solution.

Appropriate loss function

Estimation loss for model index j, the values of the parameters under both models and observation x:

$$L(\delta, (j, \theta_j, \theta_{-j})) = \mathbb{I}_{\delta=1}\mathbb{I}_{f_2(x|\theta_2)>f_1(x|\theta_1)} + \mathbb{I}_{\delta=2}\mathbf{I}_{f_2(x|\theta_2)<f_1(x|\theta_1)}$$

($\delta = j$ means model j is chosen, and $f_j(.|\theta_j)$ denotes likelihood under model j)

Under this loss, Bayes (optimal) solution

$$\delta^{\pi}(x) = \begin{cases} 1 & \text{if } \Pr^{\pi}(f_2(x|\theta_2) < f_1(x|\theta_1)|x) > \frac{1}{2} \\ 2 & \text{otherwise,} \end{cases}$$

depends on *joint* posterior distribution on (θ_1, θ_2), thus differs from Aitkin's solution.

Asymptotic properties

If $\mathcal{M}_1$ is "true" model, then $\pi(\mathcal{M}_1|x^n) = 1 + o_p(1)$ and

$$\begin{aligned}\Pr^{\pi_1}(I^1(\theta_1) > I^2(\theta_2)|x^n, \theta_2) &= \Pr(-\mathcal{X}^2_{p_1} > I^2(\theta_2) - I^2(\hat{\theta_1})) + O_p(1/\sqrt{n}) \\ &= F_{p_1}(I^1(\hat{\theta_1}) - I^2(\theta_2)) + O_p(1/\sqrt{n})\,,\end{aligned}$$

with p_1 dimension of Θ_1, $\hat{\theta}_1$ maximum likelihood estimator of θ_1
Since $I^2(\theta_2) \leqslant I^2(\hat{\theta}_2)$,

$$I^1(\hat{\theta}_1) - I^2(\theta_2) \geqslant n\mathrm{KL}(f_0, f_{\theta_2^*}) + O_p(\sqrt{n})\,,$$

where $\mathrm{KL}(f, g)$ Kullback-Leibler divergence and
$\theta_2^* = \mathrm{argmin}_{\theta_2} KL(f_0, f_{\theta_2})$, we have

$$\Pr^{\pi}(f(x^n|\theta_2) < f(x^n|\theta_1)|x^n) = 1 + o_p(1)\,.$$

Aitkin's approach leads to

$$\Pr[\mathcal{X}^2_{p_2} - \mathcal{X}^2_{p_1} > I^2(\hat{\theta}_2) - I^1(\hat{\theta}_1)],$$

thus depends on the asymptotic behavior of the likelihood ratio

[Gelman, Robert & Rousseau, 2012]

Asymptotic properties

If $\mathcal{M}_1$ is "true" model, then $\pi(\mathcal{M}_1|x^n) = 1 + o_p(1)$ and

$$\begin{aligned}\Pr^{\pi_1}(I^1(\theta_1) > I^2(\theta_2)|x^n, \theta_2) &= \Pr(-\mathcal{X}^2_{p_1} > I^2(\theta_2) - I^2(\hat{\theta_1})) + O_p(1/\sqrt{n}) \\ &= F_{p_1}(I^1(\hat{\theta_1}) - I^2(\theta_2)) + O_p(1/\sqrt{n})\,,\end{aligned}$$

with p_1 dimension of Θ_1, $\hat{\theta}_1$ maximum likelihood estimator of θ_1
Since $I^2(\theta_2) \leqslant I^2(\hat{\theta}_2)$,

$$I^1(\hat{\theta}_1) - I^2(\theta_2) \geqslant n\mathrm{KL}(f_0, f_{\theta_2^*}) + O_p(\sqrt{n})\,,$$

where $\mathrm{KL}(f, g)$ Kullback-Leibler divergence and
$\theta_2^* = \mathrm{argmin}_{\theta_2} KL(f_0, f_{\theta_2})$, we have

$$\Pr^{\pi}(f(x^n|\theta_2) < f(x^n|\theta_1)|x^n) = 1 + o_p(1)\,.$$

Aitkin's approach leads to

$$\Pr[\mathcal{X}^2_{p_2} - \mathcal{X}^2_{p_1} > I^2(\hat{\theta}_2) - I^1(\hat{\theta}_1)],$$

thus depends on the asymptotic behavior of the likelihood ratio

[Gelman, Robert & Rousseau, 2012]

uniformly most powerful "Bayesian" tests

Significance tests: one new parameter

Jeffreys-Lindley paradox

Deviance (information criterion)

Aitkin's integrated likelihood

Johnson's uniformly most powerful Bayesian tests

- AoS version
- PNAS version

Posterior predictive checking

Uniformly most powerful tests

> *"The difficulty in constructing a Bayesian hypothesis test arises from the requirement to specify an alternative hypothesis."*

Johnson's 2013 paper in the *Annals of Statistics* introduces so called *uniformly most powerful Bayesian tests*, relating to the original Neyman's and Pearson's *uniformly most powerful tests*:

$$\arg\max_{\delta} \mathbb{P}_\theta \left(\delta = 0\right), \theta \in \Theta_1$$

under the constraint

$$\mathbb{P}_\theta \left(\delta = 0\right) \leqslant \alpha, \theta \in \Theta_0$$

definition

> *"UMPBTs provide a new form of default, nonsubjective Bayesian tests in which the alternative hypothesis is determined so as to maximize the probability that a Bayes factor exceeds a specified threshold"*

i.e., find prior π_1 on Θ_1 (alternative parameter space) to maximise

$$\mathbb{P}_\theta \left(\mathfrak{B}_{10}(X) \geqslant \gamma \right) ,$$

for all $\theta \in \Theta_1$

...assuming "the null hypothesis is rejected if the posterior probability of H_1 exceeds a certain threshold"

[Johnson, 2013]

definition

> *"UMPBTs provide a new form of default, nonsubjective Bayesian tests in which the alternative hypothesis is determined so as to maximize the probability that a Bayes factor exceeds a specified threshold"*

i.e., find prior π_1 on Θ_1 (alternative parameter space) to maximise

$$\mathbb{P}_\theta \left(\mathfrak{B}_{10}(X) \geqslant \gamma \right) ,$$

for all $\theta \in \Theta_1$
...assuming "the null hypothesis is rejected if the posterior probability of H_1 exceeds a certain threshold"

[Johnson, 2013]

Examples

Example (normal mean one-sided $H_0 : \mu = \mu_0$)

H_1 point mass at

$$\mu_1 = \mu_0 + \sigma\sqrt{2\log\gamma/n}$$

and Bayes factor

$$\mathfrak{B}_{10}(z) = \exp\{z\sqrt{2\log\gamma} - \log\gamma\}$$

[Johnson, PNAS, 2013]

Examples

"Up to a constant factor that arises from the uniform distribution on μ..."

Example (normal mean two-sided $H_0 : \delta_\mu = 0$)

H_1 point mass at

$$\delta_\mu = \sigma\sqrt{2(n_1 + n_2)\log\gamma / n_1 n_2}$$

and Bayes factor

$$\mathfrak{B}_{10}(z) = \exp\{z\sqrt{2\log\gamma} - \log\gamma\}$$

[Johnson, PNAS, 2013]

Examples

Example (non-central chi-square $H_0 : \lambda = 0$)

H_1 point mass at λ^* minimum of

$$\frac{1}{\sqrt{\lambda}} \log\left(e^{\lambda/2}\gamma + \sqrt{e^{\lambda}\gamma^2 - 1}\right)$$

and Bayes factor

$$\mathfrak{B}_{10}(x) = \exp\{-\lambda^*/2\}\cosh(\sqrt{\lambda^* x})$$

[Johnson, PNAS, 2013]

Examples

Example (binomial probability one-sided $H_0 : p = p_0$)

H_1 point mass at p^* minimum of

$$\frac{\log\gamma - n[\log(1-p) - \log(1-p_0)]}{\log[p/(1-p)] - \log[p_0/(1-p_0)]}$$

and Bayes factor

$$\mathfrak{B}_{10}(x) = (p^*/p_0)^x((1-p^*)/(1-p_0))^{n-x}$$

[Johnson, PNAS, 2013]

Criticisms

- means selecting the least favourable prior under H_1 so that frequentist probability of exceeding a threshold is uniformly maximal, in a minimax perspective
- requires frequentist averaging over all possible values of the observation (violates the Likelihood Principle)
- compares probabilities for all values of the parameter θ rather than integrating against a prior or posterior
- selects a prior under H_1 with sole purpose of favouring the alternative, meaning it has no further use when H_0 is rejected
- caters to non-Bayesian approaches: Bayesian tools as supplementing p-values
- argues the method is objective because it satisfies a frequentist coverage
- very rarely exists, apart from one-dimensional exponential families
- extensions lead to data-dependent local alternatives

An impossibility theorem?

"Unfortunately, subjective Bayesian testing procedures have not been–and will likely never be–generally accepted by the scientific community. In most testing problems, the range of scientific opinion regarding the magnitude of violations from a standard theory is simply too large to make the report of a single, subjective Bayes factor worthwhile. Furthermore, scientific journals have demonstrated an unwillingness to replace the report of a single p-value with a range of subjectively determined Bayes factors or posterior model probabilities."

[Bye, everyone!]

Criticisms (2)

- use of alien notion of a "true" prior density (p.6) that would be misspecified, corresponding to "a point mass concentrated on the true value" for frequentists and to the summary of prior information for Bayesians, "not available".
- why compare probability of rejection of H_0 in favour of H_1 for *every value* of θ when (a) a prior on H_1 is used to define the Bayes factor, (b) conditioning on the data is lost, (c) the boundary or threshold γ is fixed, and (d) induced order is incomplete
- prior behind UMPB tests quite likely to be atomic, while natural dominating measure is Lebesgue
- those tests are not [NP] uniformly most powerful unless one picks a new definition of UMP tests.
- strange asymptotics: under the null

$$\log(\mathfrak{B}_{10}(X_{1:n})) \approx \mathcal{N}(-\log\gamma, 2\log\gamma)$$

goodness-of-fit?

> *"...the tangible consequence of a Bayesian hypothesis test is often the rejection of one hypothesis in favor of the second (...) It is therefore of some practical interest to determine alternative hypotheses that maximize the probability that the Bayes factor from a test exceeds a specified threshold".*

The definition of the alternative hypothesis is paramount: replacing genuine alternative H_1 with one spawned by the null H_0 voids the appeal of **B** approach, turning testing into a goodness-of-fit assessment

goodness-of-fit?

The definition of the alternative hypothesis is paramount: replacing genuine alternative H_1 with one spawned by the null H_0 voids the appeal of **B** approach, turning testing into a goodness-of-fit assessment

why would we look for the alternative that is most against H_0? See Spanos' (2013) objection of many alternative values of θ more likely than the null. This does not make them of particular interest or bound to support an alternative prior...

which threshold?

"The posterior probability of the null hypothesis does not converge to 1 as the sample size grows. The null hypothesis is never fully accepted–nor the alternative rejected–when the evidence threshold is held constant as n increases."

- notion of abstract and fixed threshold γ linked with Jeffreys-Lindley paradox
- assuming a golden number like 3 (𝔟) is no less arbitrary than using 0.05 or 5σ as significance bound (𝔣)
- even NP perspective on tests relies on decreasing (in n) Type I error types of error decreasing with n
- *in fine*, γ determined by inverting classical bound 0.05 or 0.005

which threshold?

The "behavior of UMPBTs with fixed evidence thresholds is similar to the Jeffreys-Lindley paradox"

Aspect jeopardises whole construct of UMPB tests, which depend on an arbitrary γ, unconnected with a loss function and orthogonal to any prior information

O'Bayes, anyone?

> *"...defining a Bayes factor requires the specification of both a null hypothesis and an alternative hypothesis, and in many circumstances there is no objective mechanism for defining an alternative hypothesis. The definition of the alternative hypothesis therefore involves an element of subjectivity, and it is for this reason that scientists generally eschew the Bayesian approach toward hypothesis testing.*

© Notion that is purely frequentist, using Bayes factors as the statistic instead of another divergence statistic, with no objective Bayes features and no added value

O'Bayes, anyone?

> *"The simultaneous report of default Bayes factors and p-values may play a pivotal role in dispelling the perception held by many scientists that a p-value of* 0.05 *corresponds to "significant" evidence against the null hypothesis (...) the report of Bayes factors based upon [UMPBTs] may lead to more realistic interpretations of evidence obtained from scientific studies."*

© Notion that is purely frequentist, using Bayes factors as the statistic instead of another divergence statistic, with no objective Bayes features and no added value

PNAS paper

> *"To correct this [lack of reproducibility] problem, evidence thresholds required for the declaration of a significant finding should be increased to 25–50:1, and to 100–200:1 for the declaration of a highly significant finding."*

Johnson's (2013b) recycled UMPB tests received much attention from the media for its simplistic message: move from the 0.05 significance bound to the 0.005 bound and hence reduce the non-reproducible research outcome

[Johnson, 2013b]

new arguments

- default Bayesian procedures
- rejection regions can be matched to classical rejection regions
- provide evidence in "favor of both true null and true alternative hypotheses"
- "provides insight into the amount of evidence required to reject a null hypothesis"
- adopt level 0.005 as "P values of 0.005 correspond to Bayes factors around 50"

new criticisms

- dodges the essential nature of any such automated rule, that it expresses a tradeoff between the risks of publishing misleading results and of important results being left unpublished. Such decisions should depend on costs, benefits, and probabilities of all outcomes.
- minimax alternative prior not intended to correspond to any distribution of effect sizes, solely worst-case scenario not accounting for a balance between two different losses
- threshold chosen relative to conventional value, e.g. Jeffreys' target Bayes factor of $1/25$ or $1/50$, for which there is no particular justification
- had Fisher chosen $p = 0.005$, Johnson could have argued about its failure to correspond to 200:1 evidence against the null! This $\gamma = 0.005$ turns into $z = \sqrt{-2\log(0.005)} = 3.86$, and a (one-sided) tail probability of $\Phi(-3.86) \approx 0.0005$, with no better or worse justification

[Gelman & Robert, 2013]

Posterior predictive checking

Significance tests: one new parameter

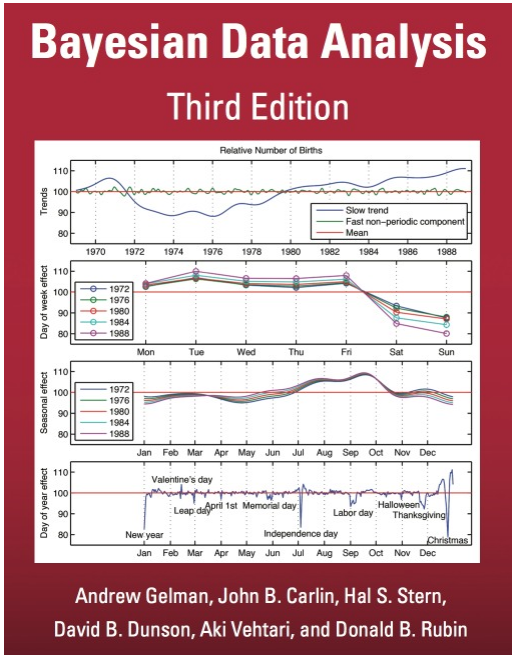
Jeffreys-Lindley paradox

Deviance (information criterion)

Aitkin's integrated likelihood

Johnson's uniformly most powerful Bayesian tests

Posterior predictive checking



Bayesian predictive

"If the model fits, then replicated data generated under the model should look similar to observed data. To put it another way, the observed data should look plausible under the posterior predictive distribution. This is really a self-consistency check: an observed discrepancy can be due to model misfit or chance." (BDA, p.143)

Use of posterior predictive

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)\pi(\theta|y)\,\mathrm{d}\theta$$

and measure of discrepancy $T(\cdot,\cdot)$
Replacing p-value

$$p(y|\theta) = \mathbb{P}(T(y^{\text{rep}},\theta) \geqslant T(y,\theta)|\theta)$$

with Bayesian posterior p-value

$$\mathbb{P}(T(y^{\text{rep}},\theta) \geqslant T(y,\theta)|y) = \int p(y|\theta)\pi(\theta|\mathbf{x})\,\mathrm{d}\theta$$

Bayesian predictive

"If the model fits, then replicated data generated under the model should look similar to observed data. To put it another way, the observed data should look plausible under the posterior predictive distribution. This is really a self-consistency check: an observed discrepancy can be due to model misfit or chance." (BDA, p.143)

Use of posterior predictive

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)\pi(\theta|y)\,\mathrm{d}\theta$$

and measure of discrepancy $T(\cdot,\cdot)$
Replacing p-value

$$p(y|\theta) = \mathbb{P}(\,T(y^{\text{rep}},\theta) \geqslant T(y,\theta)|\theta)$$

with Bayesian posterior p-value

$$\mathbb{P}(\,T(y^{\text{rep}},\theta) \geqslant T(y,\theta)|y) = \int p(y|\theta)\pi(\theta|\mathbf{x})\,\mathrm{d}\theta$$

Issues

"the posterior predictive p-value is such a [Bayesian] probability statement, conditional on the model and data, about what might be expected in future replications. (BDA, p.151)

- sounds too much like a p-value...!
- relies on choice of $T(\cdot, \cdot)$
- seems to favour overfitting
- (again) using the data twice (once for the posterior and twice in the p-value)
- needs to be calibrated (back to 0.05?)
- general difficulty in interpreting
- where is the penalty for model complexity?

Example

Normal-normal mean model:
$X \sim \mathcal{N}(\theta, 1)\,,\ \ \theta \sim \mathcal{N}(0, 10)$

Bayesian posterior p-value for
$T(x) = x^2, m(x), \mathfrak{B}_{10}(x)^{-1}$

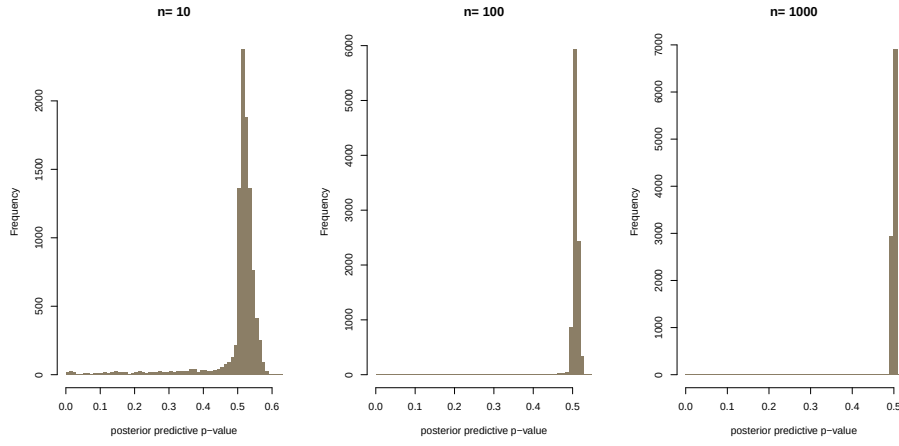
$$\int \mathbb{P}(|X| \geqslant |x||\theta, \mathbf{x})\pi(\theta|\mathbf{x})\,\mathrm{d}\theta$$

Example

Normal-normal mean model:
$X \sim \mathcal{N}(\theta, 1)\,, \;\; \theta \sim \mathcal{N}(0, 10)$

Bayesian posterior p-value for
$T(x) = x^2, m(x), \mathfrak{B}_{10}(x)^{-1}$

$$\int \mathbb{P}(|X| \geqslant |x||\theta, \mathbf{x})\pi(\theta|\mathbf{x})\, \mathrm{d}\theta$$



which interpretation?

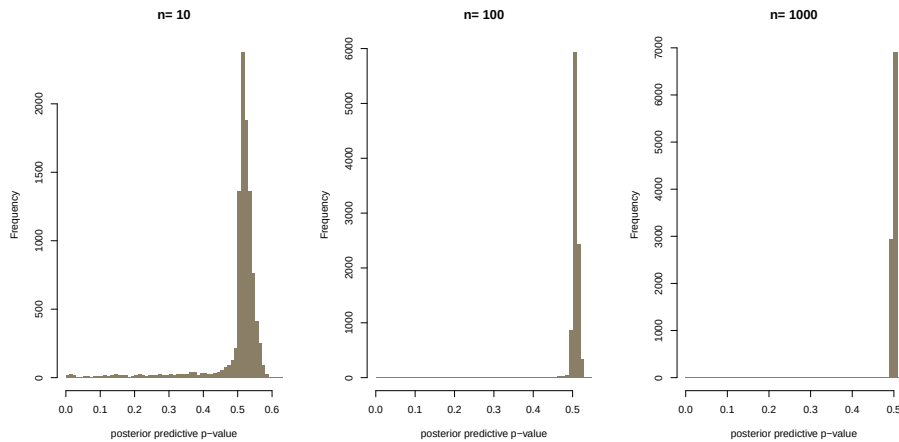
Example

Normal-normal mean model:
$X \sim \mathcal{N}(\theta, 1)\,,\ \ \theta \sim \mathcal{N}(0, 10)$

Bayesian posterior p-value for
$T(x) = x^2, m(x), \mathfrak{B}_{10}(x)^{-1}$

$$\int \mathbb{P}(|X| \geqslant |x||\theta, \mathbf{x})\pi(\theta|\mathbf{x})\,\mathrm{d}\theta$$

gets down as x gets away from 0...
while discrepancy based on $\mathfrak{B}_{10}(x)$
increases mildly



goodness-of-fit [only?]

"A model is suspect if a discrepancy is of practical importance and its observed value has a tail-area probability near 0 *or* 1*, indicating that the observed pattern would be unlikely to be seen in replications of the data if the model were true. An extreme p-value implies that the model cannot be expected to capture this aspect of the data. A p-value is a posterior probability and can therefore be interpreted directly—although not as Pr(model is true — data). Major failures of the model (...) can be addressed by expanding the model appropriately." BDA, p.150*

- not helpful in comparing models (both may be deficient)
- anti-Ockham? i.e., may favour larger dimensions (if prior concentrated enough)
- lingering worries about using the data twice and favourable bias
- impact of the prior (only under the current model) but allows for improper priors

goodness-of-fit [only?]

> *"A model is suspect if a discrepancy is of practical importance and its observed value has a tail-area probability near* 0 *or* 1*, indicating that the observed pattern would be unlikely to be seen in replications of the data if the model were true. An extreme p-value implies that the model cannot be expected to capture this aspect of the data. A p-value is a posterior probability and can therefore be interpreted directly—although not as Pr(model is true — data). Major failures of the model (...) can be addressed by expanding the model appropriately." BDA, p.150*

- not helpful in comparing models (both may be deficient)
- anti-Ockham? i.e., may favour larger dimensions (if prior concentrated enough)
- lingering worries about using the data twice and favourable bias
- impact of the prior (only under the current model) but allows for improper priors