Theory of Statistical Inference - Lecture VI.3
STA422 and STA2162

Michael Evans
University of Toronto
https://utstat.utoronto.ca/mikevans/sta422/sta4222026.html

2026

VI.2 Relative Belief (continued)

- probability model $(\Omega, \mathcal{A}, P)$ where $0 < P(A) < 1, 0 < P(C) < 1$

> **Principle of Evidence**: *If $P(A \mid C) > P(A)$, then there is evidence in favor of A being true, if $P(A \mid C) < P(A)$, then there is evidence against A being true, and if $P(A \mid C) = P(A)$, then there no evidence either way.*

- numerical measures of evidence needed as part of measuring strength of evidence (must satisfy principle of evidence with some cut-off)

- two candidates (both with cut-off 1)

$$\text{relative belief ratio } RB(A \mid C) = \frac{P(A \mid C)}{P(A)}$$

$$\text{Bayes factor } BF(A \mid C) = \frac{P(A \mid C)/P(A^c \mid C)}{P(A)/P(A^c)} = \frac{RB(A \mid C)}{RB(A^c \mid C)}$$

- which is preferred?

RBR vs BF

- how is the Bayes factor used for inference when inference base is $I^{Bayes} = (\{m_\psi : \psi \in \Psi(\Theta)\}, \pi_\Psi, x)$ for $\psi = \Psi(\theta)$?

- Bayes factor is only used for hypothesis assessment

- consider the continuous case where problem is to assess $H_0 = \{\theta : \Psi(\theta) = \psi_0\} = \Psi^{-1}\{\psi_0\}$ and $\Pi_\Psi(\{\psi_0\}) = 0$

- with relative belief we compute

$$RB_\Psi(\psi_0 \mid x) = \frac{\pi_\Psi(\psi_0 \mid x)}{\pi_\Psi(\psi_0)} = \frac{m_{\psi_0}(x)}{m(x)}$$

which tells us whether we have evidence in favor of H_0 or against H_0 as well as its strength via

$$Str_\Psi(\psi_0 \mid x) = \Pi_\Psi(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x)$$

and recall this is based on a shrinking sequence of neighbourhoods $B_\epsilon(\psi_0)$ as $\epsilon \downarrow 0$ (and in practice we discretize using $\delta =$ the difference that matters)

- if we define the Bayes factor here via such a limit we get

$$BF(\psi_0 \mid x) = \lim_{\epsilon \downarrow 0} BF(B_\epsilon(\psi_0) \mid x) = RB_\Psi(\psi_0 \mid x)$$

- but in this context the Bayes factor is not defined as a limit, rather a prior π_{ψ_0} is placed on $\Psi^{-1}\{\psi_0\}$, a positive mass p_0 is placed on $\Psi^{-1}\{\psi_0\}$ and the mixture (spike-and-slab) prior

$$\pi_{p_0,\psi_0}(\theta) = p_0 \pi_{\psi_0}(\theta) + (1 - p_0)\pi(\theta)$$

is used and note now the prior probability of H_0 is p_0 and this leads to, where $m_{p_0,\psi_0}(x) = \int_{\Psi^{-1}\{\psi_0\}} f_\theta(x)\pi_{\psi_0}(\theta)\, d\theta$,

$$BF_{p_0,\psi_0}(\psi_0 \mid x) = \frac{m_{p_0,\psi_0}(x)}{m(x)}$$

- but note, π induces a conditional probability distribution $\pi(\theta \mid \psi_0)$ on $\Psi^{-1}\{\psi_0\}$ as a limit (prior on nuisance parameters)

- so if $\pi_{\psi_0} \neq \pi(\cdot \mid \psi_0)$ there is a contradiction in the expression of beliefs and when $\pi_{\psi_0} = \pi(\cdot \mid \psi_0)$, then $BF_{p_0,\psi_0}(\psi_0 \mid x) = RB_\Psi(\psi_0 \mid x)$

- so there is a fundamental problem with the definition of the BF if we don't take $\pi_{\psi_0} = \pi(\cdot \mid \psi_0)$ and then you wonder why bother with defining/using the BF

- the strength of the evidence given by the Bayes factor is measured via Jeffreys' scale

BF	Strength
1 to $10^{1/2}$	Barely worth mentioning
$10^{1/2}$ to 10	Substantial
10 to $10^{3/2}$	Strong
$10^{3/2}$ to 10^2	Very Strong
$> 10^2$	Decisive

Example VI.2.3 (*Jeffreys-Lindley/Bartlett paradox*)

- $\bar{x} \sim N(\mu, \sigma_0^2/n)$ and interest is in the hypothesis $H_0 = \{0\}$

- suppose $T_{\mu_0}(x) = \sqrt{n}|\bar{x}|/\sigma_0 = 5$ is obtained which leads to the p-value $2(1 - \Phi(\sqrt{n}|\bar{x}|/\sigma_0)) = 5.733031 \times 10^{-07}$

- suppose the prior $\mu \sim N(0, \tau_0^2)$ is used

- note that $RB(0 \mid x) = BF_{p_0,\psi_0}(0 \mid x)$ here because the only prior on $\{0\}$ is the degenerate one and

$$RB(0 \mid x) = \sqrt{1 + n\tau_0^2} \exp\left\{ -\frac{n\bar{x}^2}{2} \frac{n\tau_0^2}{1 + n\tau_0^2} \right\}$$

- Jeffreys-Lindley - consider $\sqrt{n}\bar{x} = 5$ fixed and let $n \to \infty$, then $RB(0 \mid x) \to \infty$ which indicates evidence in favor and by Jeffreys scale decisive evidence in favor

- Bartlett - fix n and let $\tau_0^2 \to \infty$, so statistician has little information about the true value, then again $RB(0 \mid x) \to \infty$

- but then the strength is given by

$$\begin{aligned}
&\Pi(\,RB(\mu \,|\, x) \leq RB(0 \,|\, x) \,|\, x) \\
&= 1 - \Phi((1 + 1/n\tau_0^2)^{1/2}(|\sqrt{n}\bar{x}| + (n\tau_0^2 + 1)^{-1}\sqrt{n}\bar{x})) + \\
&\Phi((1 + 1/n\tau_0^2)^{1/2}(-|\sqrt{n}\bar{x}| + (n\tau_0^2 + 1)^{-1}\sqrt{n}\bar{x}))
\end{aligned}$$

and for both contexts

$$\begin{aligned}
Str_\Psi(0 \,|\, x) &= \Pi(\,RB(\mu \,|\, x) \leq RB(0 \,|\, x) \,|\, x) \rightarrow 2(1 - \Phi(|\sqrt{n}\bar{x}|) \\
&= 5.733031 \times 10^{-07}
\end{aligned}$$

which indicates very weak evidence in favor ■

- does this resolve the problem? No, we need the concept of bias for that

Bias calculations for H

- we ask the question: given the model and prior and the amount of data we will collect, is there bias in favor of or against H_0?

- how to measure this?

- by the following prior probabilities

$$\textit{bias against}_{\Psi}(\psi_0) = M_{\psi_0}(RB_{\Psi}(\psi_0 \mid X) \leq 1)$$

$=$ prior probability of not getting evidence in favor of H_0 when it is true,

$$\textit{bias in favor of}_{\Psi}(\psi_0) = \sup_{\psi : d_{\Psi}(\psi, \psi_0) \geq \delta} M_{\psi}(RB_{\Psi}(\psi_0 \mid X) \geq 1)$$

$=$ prior probability of getting evidence in favor of H_0
when it is meaningfully false

- we want these probabilities to be small to say inferences drawn are reliable

- a prior that minimizes one bias maximizes the other bias but both biases $\longrightarrow 0$ as the amount of data increases

Example VI.2.3 (*Jeffreys-Lindley/Bartlett paradox continued*)

- *bias against*$(\mu_0) \longrightarrow 0$ and the *bias in favor of* $(\mu_0) \longrightarrow 1$ in both scenarios

- Resolution: don't choose the prior to be arbitrarily diffuse to reflect noninformativeness, rather choose a prior that is sufficiently diffuse to cover the interval where it is known μ must lie, e.g., the interval where the measurements lie, and then choose n so that both biases are suitably small ■

- note - the bias calculations are frequentist with respect to the model $\{m_\psi : \psi \in \Psi(\Theta)\}$

- it seems reasonably clear that the Bayes factor defined by the spike-and-slab prior is not correct and even without that prior, Jeffreys scale does not the measure the strength of evidence properly

- whatever measure of evidence is used, it needs to be calibrated: **there is no universal scale on which evidence is measured**

Bias calculations for E

- another problem with the BF is that it can't be used for the **E** problem, at least as currently employed, in the continuous case

- with relative belief the natural estimate is $\psi(x) =$ MLE under the model $\{m_\psi : \psi \in \Psi(\Theta)\}$

- bias for this problem is defined in terms of coverage probabilities for plausible and implausible regions

$$\begin{aligned} Pl_\Psi(x) &= \{\psi : RB_\Psi(\psi \mid x) > 1\} \\ &= \text{ values with evidence in favor} \\ Im_\Psi(x) &= \{\psi : RB_\Psi(\psi \mid x) < 1\} \\ &= \text{ values with evidence against} \end{aligned}$$

- then biases are given by, where d_Ψ is a distance measure on $\Psi(\Theta)$

$$
\begin{aligned}
\textit{bias against}_\Psi &= E_{\Pi_\Psi}(M_\psi(RB_\Psi(\psi \mid X) \leq 1)) = E_{\Pi_\Psi}\left(M_\psi(\psi \notin Pl_\Psi(X))\right) \\
&= \text{ prior probability that true value is not in } Pl_\Psi(x), \\
\textit{bias in favor of}_\Psi &= E_{\Pi_\Psi}\Big(\sup_{\psi' : d_\Psi(\psi', \psi) \geq \delta} M_{\psi'}(RB_\Psi(\psi \mid X) \geq 1)\Big) \\
&= E_{\Pi_\Psi}\left(\sup_{\psi' : d_\Psi(\psi', \psi) \geq \delta} M_{\psi'}(\psi \notin Im_\Psi(X))\right), \\
&= \text{ prior probability that a meaningfully false value} \\
&\qquad \text{is in } Pl_\Psi(x)
\end{aligned}
$$

- note $1 - E_{\Pi_\Psi}\left(M(\psi \notin Pl_\Psi(X) \mid \psi)\right)$ is the prior coverage probability (*Bayesian confidence*) of $Pl_\Psi(x)$

- there is typically a value $\psi_* = \arg\sup_\psi \mathit{bias\ against}_\Psi(\psi)$ and so

$$M_\psi(\psi \in Pl_\Psi(X)) \geq 1 - M_{\psi^*}(RB_\Psi(\psi_* \mid X) \leq 1)$$

gives a lower bound on the frequentist confidence of $Pl_\Psi(x)$ with respect to the model $\{m(\cdot \mid \psi) : \psi \in \Psi\}$

- note - evidence for or against, the plausible region and the bias calculations are completely independent of the valid measure of evidence chosen

Example VI.2.4 *binomial*

- $n\bar{x} \sim$ binomial(n, θ) and we want to estimate $\theta \in [0, 1]$

- prior $\theta \sim$ beta(α, β) gives posterior beta$(n\bar{x} + \alpha, n(1 - \bar{x})\beta)$ and

$$RB(\theta \mid \bar{x}) = \frac{\binom{n}{n\bar{x}}\theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}}{m(x)}$$

where

$$m(\bar{x}) = \binom{n}{n\bar{x}} \frac{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)}{\Gamma(n+\alpha+\beta)}$$

- relative belief estimate is $\theta(x) = \bar{x}$ and

$$\begin{aligned} PI(\bar{x}) &= \{\theta : RB(\theta \mid \bar{x}) > 1\} \\ &= \left\{\theta : \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})} > \frac{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)}{\Gamma(n+\alpha+\beta)}\right\} \end{aligned}$$

and note that this is a likelihood interval

- to obtain the biases we need to proceed via simulation

- here is an outline of a simulation to determine the confidence of $Pl(\bar{x})$

  1. *choose a grid of* N *values of* $\theta \in [0,1]$
  2. *for each* θ *in the grid generate* $n\bar{x} \sim$ *binomial*(n,θ), *compute* $RB(\theta \mid \bar{x})$ *and determine if* $RB(\theta \mid \bar{x}) \leq 1$, *repeat multiple times and estimate* $M_\theta(RB(\theta \mid X) \leq 1)$
  3. *find value* $\theta^* = \arg\sup M_\theta(RB(\theta \mid X) \leq 1)$, *then* $1 - M_{\theta^*}(RB(\theta^* \mid X) \leq 1)$ *is the confidence level of* $Pl(\bar{x})$

- if the coverage of $Pl(\bar{x})$ is not high enough, then increase n

- prior given by $(\alpha, \beta) = (1, 1)$

n	max bias against	confidence	bias against	Bayes confidence
10	0.21	0.79	0.11	0.89
50	0.07	0.93	0.05	0.95
100	0.05	0.95	0.03	0.97

- prior given by $(\alpha, \beta) = (5, 5)$

n	max bias against	confidence	bias against	Bayes confidence
10	0.36	0.64	0.21	0.79
50	0.16	0.84	0.10	0.90
100	0.11	0.89	0.07	0.93

- prior given by $(\alpha, \beta) = (1, 1), \delta = 0.1$

n	bias in favor
10	0.84
50	0.51
100	0.35

■

- relative belief inferences have a number of nice properties

- if we reparameterize from ψ to $\lambda = \Lambda(\psi)$ via 1-1 smooth Λ, then

$$\begin{aligned} RB_{\Lambda}(\lambda \mid x) &= \frac{\pi_{\Lambda}(\lambda \mid x)}{\pi_{\Lambda}(\lambda)} = \frac{\pi_{\Psi}(\Lambda^{-1}(\lambda) \mid x) J_{\Lambda} \Lambda^{-1}(\lambda)}{\pi_{\Psi}(\Lambda^{-1}(\lambda)) \Lambda^{-1}(\lambda)} \\ &= \frac{\pi_{\Psi}(\psi \mid x)}{\pi_{\Psi}(\psi)} = RB_{\Psi}(\psi \mid x) \end{aligned}$$

so all relative belief inferences are invariant under reparameterizations

- relative belief inferences are optimally robust, among all Bayesian inferences, to the prior π_{Ψ}

> *Al-Labadi and Evans (2017) Optimal robustness results for some Bayesian procedures and the relationship to prior-data conflict. Bayesian Analysis 12, 3, 702-728.*

- more on relative belief ratios versus Bayes factors

> *Al-Labadi, Alzaatreh and Evans (2025) How to measure evidence and its strength: Bayes factors or relative belief ratios? Canadian Journal of Statistics, 53, 4.*

- the prior probability that a credible region $B_{\Psi,\gamma}(x)$ covers a false value is

$$E_M(\Pi_\Psi(B_{\Psi,\gamma}(x))) = \int_\times \int_{\Psi(\Theta)} P_\theta(\psi \in B_{\Psi,\gamma}(x))\,\Pi_\Psi(d\psi)\,\Pi(d\theta)$$

then if $\Pi_\Psi(C_{\Psi,\gamma}(x)\,|\,x) = \gamma$ for every x, then
$E_M(\Pi_\Psi(B_{\Psi,\gamma}(x))) \geq E_M(\Pi_\Psi(C_{\Psi,\gamma}(x)))$

- the prior probability of covering the true value is

$$E_M(\Pi_\Psi(C_{\Psi,\gamma}(x)\,|\,x)) = \int_\times P_\theta(\Psi(\theta) \in C_{\Psi,\gamma}(x))\,\Pi(d\theta)$$

and we always have

$$E_M(\Pi_\Psi(C_{\Psi,\gamma}(x)\,|\,x)) \geq E_M(\Pi_\Psi(C_{\Psi,\gamma}(x)))$$

so a γ-relative belief region is always unbiased

- etc.

- elicitation

Example VI.2.5 (*location normal*)

- $\bar{x} \sim N(\mu, \sigma_0^2/n)$ with $\mu \in \mathbb{R}$ unknown

- a prior on μ should reflect what is known

- the data arises via a measurement process

- so we know that the data will fall in some interval (l, u)

- choose a convenient prior that reflects this information, such as $\mu \sim N(\mu_0, \tau_0^2)$

- this is a conjugate prior because the posterior is of the same form as the prior, namely,

$$\mu \,|\, \bar{x} \sim N\left(\left(\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)^{-1}\left(\frac{n\bar{x}}{\sigma_0^2} + \frac{\mu_0}{\tau_0^2}\right), \left(\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)^{-1}\right)$$

- reasonable choice $\mu_0 = (l+u)/2$ and then choose τ_0^2 so that the prior probability content of the interval (l,u) is some large probability (virtual certainty) γ so want

$$\begin{aligned}\gamma &= \Phi\left(\frac{u-\mu_0}{\tau_0}\right) - \Phi\left(\frac{l-\mu_0}{\tau_0}\right) = \Phi\left(\frac{u-l}{2\tau_0}\right) - \Phi\left(\frac{l-u}{2\tau_0}\right) \\ &= 2\Phi\left(\frac{u-l}{2\tau_0}\right) - 1 \text{ which implies } \tau_0 = \frac{(u-l)/2}{z_{(1+\gamma)/2}} \blacksquare\end{aligned}$$

- currently there is no place for improper priors in relative belief because they don't express beliefs and arguments for using such priors don't seem overly convincing

- given that an improper prior is not a probability distribution, what is called the posterior that arises in such contexts, is not a conditional probability distribution

- similarly empirical Bayes approaches, where a prior is chosen in a family $\{\pi_\lambda : \lambda \in \Lambda\}$ based on the observed data x, say producing $\pi_{\lambda(x)}$, do not produce posteriors via conditioning because $\pi_{\lambda(x)}(\theta) f_\theta(x)$ cannot be considered as the joint distribution of (θ, x)

- the inference base is $I^{Bayes} = (\{m_\psi : \psi \in \Psi(\Theta)\}, \pi_\Psi, x)$ where the model $\{m_\psi : \psi \in \Psi(\Theta)\}$ and prior π_Ψ

- joint for (ψ, x), when T is a mss for $\{m_\psi : \psi \in \Psi(\Theta)\}$

$$\begin{aligned} \pi_\Psi(\psi) m_\psi(x) &= \pi_\Psi(\psi) m_{\psi,T}(T(x)) m(x \mid T(x)) \\ &= \pi_\Psi(\psi \mid T(x)) m_T(T(x)) m(x \mid T(x)) \end{aligned}$$

- logically it makes sense to check the model first (if the model is substantially wrong, then the prior is irrelevant)

- $m(x \mid T(x))$ is available for checking: is the data x surprising in terms of $m(x \mid T(x))$, e.g., measure the location of x wrt this distribution

$$M_T(m(X \mid T(x)) \leq m(x \mid T(x)) \mid T(x))$$

and when this is small this would indicate a problem

- if the model seems fine, then $m_T(T(x))$ is available for checking the prior via

$$M_T(m_T(t) \leq m_T(T(x)))$$

which locates the value of $T(x)$ in its marginal distribution and if it is small, then this is prior-data conflict

- why? under conditions, as the amount of data increases

$$M_T(m_T(t) \leq m_T(T(x))) \to \Pi_\Psi\left(\pi_\Psi(\psi) \leq \pi_\Psi(\psi_{true})\right)$$

Evans, M. and Jang, G-H. (2011) A limit result for the prior predictive applied to checking for prior-data conflict. Statistics and Probability Letters, 81, 1034-1038.

- note that this does not suffer form the large sample paradox as increasing the amount of data will not necessarily lead to a conclusion that a prior-data conflict exists

- if ancillary statistics exist, these can also be part of model checking and checking for prior-data conflict see

Evans, M. and Moshonov, H. (2006) Checking for prior-data conflict. Bayesian Analysis, 1, 4, 893-914.

Example VI.2.5 (*location normal continued*)

- $T(x) = \bar{x}$ and the conditional distribution $x \mid T(x)$ is given by

$$x - \bar{x}1_n \sim N_n\left(0, \sigma_0^2 \begin{pmatrix} 1-1/n & 1-2/n & \cdots \\ 1-2/n & \ddots & \vdots \\ \vdots & \cdots & 1-1/n \end{pmatrix}\right)$$

- the prior distribution of $\bar{x} \sim N(\mu_0, \sigma_0^2/n + \tau_0^2)$ so

$$M_T(m_T(t) \leq m_T(T(x))) = P(\left(\sigma_0^2/n + \tau_0^2\right)^{-1}(\bar{X} - \mu_0)^2 \leq \left(\sigma_0^2/n + \tau_0^2\right)^{-1}$$

and

$$\left(\sigma_0^2/n + \tau_0^2\right)^{-1}(\bar{X} - \mu_0)^2 \sim \text{ chi-squared}(1)$$

■

What do you do when one of the ingredients fails?

- what to do when the model fails its checks?

- there really isn't a lot that can be said about this in general, although some techniques like transforming the data have been developed

- what to do when prior-data conflict is detected?

- typically prior-data conflict arises when true value lies in the tails of the prior

- so a logical approach is to modify the prior so that this is avoided

- an organized approach to this is developed in

> *Evans and Jang (2011). Weak informativity and the information in one prior relative to another. Statistical Science, 26, 3, 423-439.*

- basically, starting from a base elicited prior, we construct a hierarchy of priors that are progressively weakly informative as one moves up the hierarchy where the degree of weak informativity is quantified and this is done before seeing the data

- when prior-data conflict is detected one moves up the hierarchy until the conflict is avoided