Theory of Statistical Inference - Lecture VI.2
STA422 and STA2162

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

2026

VI.2 Relative Belief

- a different approach to deriving Bayesian inferences based on a clear definition of what is meant by statistical evidence with inference base $I^{Bayes} = (\{f_\theta : \theta \in \Theta\}, \pi, x)$

- based on the idea that the purpose of a statistical study is to report what the evidence in the data says about the questions of interest posed in a scientific investigation

- the role of the subject of statistics is to say, without ambiguity, how such a study should proceed

- we can identify the following steps

1. Choose the ingredients: model $\{f_\theta : \theta \in \Theta\}$ and prior π (elicitation) and identify object of interest $\psi = \Psi(\theta)$.

2. Design the study based on the bias calculations: e.g., decide how much data is required to obtain reliable results which is essentially frequentist.

3. Once the data is obtained via an acceptable approach (and so is objective) check that the ingredients chosen are not contradicted by the data (model checking and checking for prior-data conflict) and, if they are, replace them.

4. Inference via relative belief.

- note - these steps (more needed?) represent an ideal which is necessary for the development of an appropriate theory of statistical reasoning

- when this cannot be attained, e.g., step 2 was not carried out, qualifications need to be stated about any inferences reported

- note that the bias calculations can be done post hoc (after seeing the data) and this will tell us about the reliability of the study

- this lecture will discuss 4 and 2 and next week 1 and 3

- note - inferences based on evidence is not in antagonism with decisions, as evidence can be ignored when decisions are made, but still inference that reflects the evidence is where the science is as it represents what has been learned

- to start, consider a probability model $(\Omega, \mathcal{A}, P)$ and events $A, C \in \mathcal{A}$ where $0 < P(A) < 1, 0 < P(C) < 1$

- suppose interest is in whether or not an unobserved $\omega \in \Omega$ satisfies $\omega \in A$ (is A true)

- the initial belief that A is true is given by $P(A)$, but suppose we are told, through a valid information generator, that $\omega \in C$ (the data)

- then the principle of conditional probability says that our initial belief $P(A)$ is replaced by $P(A \mid C)$

- we can now characterize the evidence concerning the truth of A

> **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.*

- note - $P(A \mid C) = P(A)$ iff A and C are statistically independent so nothing can be learned about A by observing that C is true

- this characterizes statistical evidence in a simple, clear, intuitively satisfying way

- this is not new (although the name may be due to me)

- Popper (1968) The Logic of Scientific Discovery (Appendix ix) writes

> *If we are asked to give a criterion of the fact that the evidence y supports or corroborates a statement x, the most obvious reply is: that y increases the probability of x.*

- within the discipline of the philosophy of science there is *confirmation theory* which is based on this principle, see Salmon,W. (1973) Confirmation. Scientific American, 228, 75–81

- there is lots of discussion about this idea in the philosophy of science literature and sometimes supposed counterexamples are presented

Example VI.2.1 *Hempel's the Raven Paradox*

- $A =$ "all ravens are black"

- truth value of A is equivalent to the truth value of the contrapositive "all nonblack objects are not ravens"

- you observe an object in this room that is not black and is not a raven and so this confirms A (our belief that A is true increases) and this seems spurious ■

- as will be discussed (when we discuss bias) this, like the many other counterexamples form philosophy, is that they are not put into a statistical framework which is very likely the proper context for any theory of induction

- this is just a bad experiment because the outcome is a foregone conclusion with high prior probability

- there is more to measuring evidence than just determining if there is evidence in favor or otherwise, we have to say how strong the evidence is

- a natural measure of strength is given by $P(A \mid C)$, because it measures how strongly we believe what the evidence indicates

- suppose we have evidence in favor of A, then a large value of $P(A \mid C)$ indicates that we have strong evidence in favor and a small value of $P(A \mid C)$ indicates only weak evidence in favor

- by contrast, if we have evidence against, then a large value of $P(A \mid C)$ indicates that we have weak evidence against and a small value of $P(A \mid C)$ indicates strong evidence against

- there are situations in Bayesian inference where this approach makes sense but there are several issues associated with it, e.g. $P(A)$ small may force $P(A \mid C)$ to be small too unless C is quite definitive ($C \subset A$ implies $P(A \mid C) = 1$)

- we need a numerical measure of evidence so we can calibrate more effectively when we have strong or weak evidence

Definition VI.2.1 A measure of evidence $ev(A \mid C)$ for the event A after observing C is true, is a *valid measure of evidence* if there is a cut-off c such that

$$\begin{aligned} ev(A \mid C) &> c \text{ iff } P(A \mid C) > P(A) \\ ev(A \mid C) &< c \text{ iff } P(A \mid C) < P(A) \\ ev(A \mid C) &= c \text{ iff } P(A \mid C) = P(A). \blacksquare \end{aligned}$$

- there are a number of valid measures of evidence but we will focus on the *relative belief ratio*

$$RB(A \mid C) = \frac{P(A \mid C)}{P(A)}$$

with cut-off 1 and any 1-1 increasing function of $RB(A \mid C)$ such as $\log RB(A \mid C)$ with cut-off 0

- a commonly used measure of evidence is the *Bayes factor*

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

which is the ratio of the posterior odds in favor of A being true to the prior odds in favor of A being true

Lemma VI.2.1 The Bayes factor is a valid measure of evidence with cut-off 1.
Proof:

$$BF(A \mid C) \quad > \quad (<,=)1 \text{ iff } \frac{P(A \mid C)}{1-P(A \mid C)} > (<,=)\frac{P(A)}{1-P(A)}$$

$$\text{iff } 1/P(A \mid C)\,(\quad > \quad ,=) < 1/P(A) \text{ iff } P(A \mid C)\,(<,=) > 1/P(A). \blacksquare$$

- there are others (e.g. $P(A \mid C) - P(A)$ with cut-off 0), but the RB has a number of good properties and the BF is commonly used

Lemma VI.2.2

(i) $P(A \mid C) \leq RB(A \mid C) \leq 1/P(A)$.

(ii) (symmetry) $RB(A \mid C) = RB(C \mid A)$ since

$$RB(C \mid A) = \frac{P(C \mid A)}{P(C)} = \frac{P(A \cap C)}{P(A)P(C)} = \frac{P(A \mid C)}{P(A)}$$

so the evidence directions are the same but note the strength measures will be different ($P(A \mid C)$ versus $P(C \mid A)$).

(iii) $RB(A^c \mid C) = \frac{1-P(A)RB(A \mid C)}{1-P(A)}$ so $RB(A \mid C) > (<, =)1$ iff $RB(A^c \mid C) < (>, =)1$.

(iv)

$$RB(A^c \mid C^c) = \frac{P(A^c \mid C^c)}{P(A^c)} = \frac{P(A^c \cap C^c)}{P(A^c)P(A^c)}$$

$$= \frac{1 - P(A \cup C)}{(1 - P(A))(1 - P(C))}$$

$$= \frac{1 - P(A) - P(C) + P(A \cap C)}{1 - P(A) - P(C) + P(A)P(C)}$$

$$= \frac{1 - P(A) - P(C) + P(A)P(C)RB(A \mid C)}{1 - P(A) - P(C) + P(A)P(C)}$$

so $RB(A \mid C) > (<, =)1$ iff $RB(A^c \mid C^c) > (<, =)1$.

(v) $BF(A \mid C) = \frac{RB(A \mid C)}{RB(A^c \mid C)}$ so we always have $BF(A \mid C) > (<) RB(A \mid C)$ when $RB(A \mid C) > (<)1$ and $RB(A \mid C) = \frac{BF(A \mid C)}{1 - P(A) + P(A)BF(A \mid C)}$ so BF can be expressed simply in terms of RB but not conversely

(vi) When A and B are conditionally independent given C

$$\begin{aligned} RB(A \cap B \mid C) &= \frac{P(A \cap B \mid C)}{P(A \cap B)} = \frac{P(A \mid C)P(B \mid C)}{P(A \mid B)P(B)} \\ &= \frac{RB(A \mid C)RB(B \mid C)}{RB(A \mid B)} \end{aligned}$$

and when A and B are also independent

$$RB(A \cap B \mid C) = RB(A \mid C)RB(B \mid C).$$

(vii) When $\{A_1, \ldots, A_k\}$ forms a partition of Ω with $P(A_i) > 0$ for $i = 1, \ldots, k$

$$\begin{aligned} 1 &= RB(\Omega \mid C) = RB(\cup_{i=1}^k A_i \mid C) = \frac{\sum_{i=1}^k P(A_i \mid C)}{\sum_{i=1}^k P(A_i)} \\ &= \sum_{i=1}^k P(A_i) RB(A_i \mid C). \end{aligned}$$

- for $A \subset B$ we can have $RB(A \mid C) > 1$ but $RB(B \mid C) < 1$, so RB is not monotonic, is this incoherent?

Lemma VI.2.3. When $A \subset B$ and all the events $A, B \backslash A$ and C have positive probability, then $RB(A \mid C) > RB(B \mid C)$ iff $RB(A \mid C) > RB(B \backslash A \mid C)$.

Proof:

$$RB(A \mid C) = \frac{P(A \mid C)}{P(A)} > RB(B \mid C) = \frac{P(B \mid C)}{P(B)} \text{ iff}$$

$$\frac{P(B)}{P(A)} = \frac{P(B \backslash A)}{P(A)} + 1 > \frac{P(B \backslash A \mid C)}{P(A \mid C)} + 1 \blacksquare$$

Example VI.2.2

- $\Omega =$ population of a country where B is the set of adults over 20 years of age and $A \subset B$ is the set of university graduates over 20

- individual ω is randomly selected and it is determined that $\omega \in C =$ those in favor of vaccination against COVID-19

- it is certainly reasonable that the proportion of university graduates over 20 in favor of vaccination, within the subpopulation of those who believe in vaccination, is larger than proportion of university graduates within the population of those over 20, namely, $P(A \mid C) > P(A)$,so observing $\omega \in C$ is evidence in favor of $\omega \in A$

- a similar comment applies to B but it would not be at all surprising, however, that $RB(A \mid C) > RB(B \backslash A \mid C)$ since $B \backslash A$ are those over 20 with less education ■

- it might even be the case that there is evidence against B, when will this occur?

Lemma VI.2.4. Under the conditions of Lemma VI.2.3, if $RB(A \mid C) > 1$, then $RB(B \mid C) < 1$ iff $RB(B \backslash A \mid C) < 1$ and

$$P(A \mid B) < \frac{1 - RB(B \backslash A \mid C)}{RB(A \mid C) - RB(B \backslash A \mid C)}. \tag{1}$$

Proof: Using $\{A, B \backslash A\}$ is a partition of B and Lemma VI.2.2 (vii)

$$RB(B \mid C) = RB(A \mid C)P(A \mid B) + RB(B \backslash A \mid C)P(B \backslash A \mid B)$$

and then $RB(B \mid C) < 1$ forces $RB(B \backslash A \mid C) < 1$ since $RB(A \mid C) > 1$ and $P(A \mid B) > 0$. Then

$$\begin{aligned} RB(A \mid C)P(A \mid B) + RB(B \backslash A \mid C)(1 - P(A \mid B)) &< 1 \text{ iff} \\ (RB(A \mid C) - RB(B \backslash A \mid C))P(A \mid B) &< 1 - RB(B \backslash A \mid C) \end{aligned}$$

which gives (1). Now suppose $RB(B \backslash A \mid C) < 1$ and (1) holds. Then from (1)

$$RB(B \mid C) = (RB(A \mid C) - RB(B \backslash A \mid C))P(A \mid B) + RB(B \backslash A \mid C) < 1. \blacksquare$$

- this occurs whenever $RB(B \backslash A \mid C) < 1$ (nonuniversity grads generally don't believe in vaccination) and either $P(A \mid B)$ is very small (proportion of population over 20 that are university grads) or $RB(A \mid C) \approx 1$

- so we have $I^{Bayes} = (\{f_\theta : \theta \in \Theta\}, \pi, x)$ and $\psi = \Psi(\theta)$

- when each f_θ and π are discrete, then having observed $C = \{x\}$ and letting $A = \{\psi\}$,

$$RB_\Psi(\psi \mid x) = \frac{\pi_\Psi(\psi \mid x)}{\pi_\Psi(\psi)}$$

- and note, using $RB(A \mid C) = RB(C \mid A)$

$$RB_\Psi(\psi \mid x) = RB_X(x \mid \psi) = \frac{m_\psi(x)}{m(x)}$$

where $m_\psi(x) = \sum_{\theta \in \Psi^{-1}\{\psi\}} \pi(\theta \mid \psi) f_\theta(x)$

- what do we do when we have a continuous prior?

- let $B_\epsilon(\psi)$ be a neighbourhood of ψ that shrinks nicely to $\{\psi\}$ as $\epsilon \downarrow 0$, then when π_Ψ is continuous and positive at ψ,

$$RB_\Psi(\psi \mid x) \stackrel{def}{=} \lim_{\epsilon \downarrow 0} \frac{\Pi_\Psi(\psi \mid x)}{\Pi_\Psi(\psi)} = \frac{\pi_\Psi(\psi \mid x)}{\pi_\Psi(\psi)}$$

and again we have

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

where $m_\psi(x) = \int_{\Psi^{-1}\{\psi\}} \pi(\theta \mid \psi) f_\theta(x)\, d\theta$

- note - when $\Psi(\theta) = \theta$, then

$$RB(\theta \mid x) = \frac{\pi(\theta \mid x)}{\pi(\theta)} = \frac{f_\theta(x)}{m(x)}$$

so the likelihood based on $(\{f_\theta : \theta \in \Theta\}, x)$ is proportional to $RB(\theta \mid x)$ although $RB(\theta \mid x)$ can't be multiplied by a constant because the cut-off of 1, determining evidence for or against, would change

- in general

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

the likelihood based on $(\{m_{\psi} : \psi \in \Psi(\Theta)\}, x)$ is proportional to $RB_{\Psi}(\psi \mid x)$

- so in the continuous case

$$RB_{\Psi}(\psi \mid x) = \frac{\pi_{\Psi}(\psi \mid x) Vol(B_{\epsilon}(\psi))}{\pi_{\Psi}(\psi) Vol(B_{\epsilon}(\psi))} \approx \frac{\text{posterior prob. of } \psi}{\text{prior prob. of } \psi}$$

for reasonably small ϵ

H

- suppose we want to assess $H_0 = \{\psi_0\}$, then

$$\begin{array}{ll} RB_\Psi(\psi_0 \mid x) > 1 & \text{evidence in favor of } H_0 \\ RB_\Psi(\psi_0 \mid x) < 1 & \text{evidence against } H_0 \\ RB_\Psi(\psi_0 \mid x) = 1 & \text{no evidence either way} \end{array} \quad > 1$$

- what about the strength of this evidence?

- could quote $\Pi_\Psi(\{\psi_0\} \mid x)$ in the discrete case but $\Pi_\Psi(\psi_0 \mid x) = 0$ in the continuous case and, even in the discrete case, $\Pi_\Psi(\{\psi_0\} \mid x)$ may be small because $\Pi_\Psi(\{\psi_0\})$ is small

- the problem here is that we need to calibrate $RB_\Psi(\psi_0 \mid x)$ and there is no universal sale for doing this so we have to compare $RB_\Psi(\psi_0 \mid x)$ to each of the other values $RB_\Psi(\psi \mid x)$ to see if it large, when there is evidence in favor or when there is evidence against, see if it is small

- one way to do this is to compute the strength

$$\begin{aligned} Str_{\Psi}(\psi_0 \,|\, x) &= \Pi_{\Psi}(RB_{\Psi}(\psi \,|\, x) \leq RB_{\Psi}(\psi_0 \,|\, x) \,|\, x) \\ &= \text{ posterior prob. of the true value of } \psi \\ &\quad \text{having a relative belief ratio no larger than that of } \psi_0 \end{aligned}$$

- so when $RB_{\Psi}(\psi_0 \,|\, x) > 1$ and $Str_{\Psi}(\psi_0 \,|\, x) \approx 1$ we have strong evidence in favor of ψ_0, because there is only small belief that the true value has a bigger RB ratio than ψ_0, and when $RB_{\Psi}(\psi_0 \,|\, x) < 1$ and $Str_{\Psi}(\psi_0 \,|\, x) \approx 0$ we have strong evidence against ψ_0, because there is large belief that the true value has a bigger RB ratio than ψ_0

- it is important that the values $\psi = \Psi(\theta)$ are similar in nature so we aren't comparing an apple with an orange

Important caveat in the case where ψ is a continuous parameter

- we need to be clear about $\delta =$ the difference that matters

- so, if d is a distance measure on $\Psi(\Theta)$ we need to be able to say that when $d(\psi_1, \psi_2) < \delta$ then ψ_1, ψ_2 are effectively the same as far as the application goes

- suppose $\psi \in \mathbb{R}$ and we want to know ψ to accuracy given by δ

- so we prescribe a grid $\{\ldots \psi_{-2}, \psi_{-1}, \psi_0, \psi_1, \psi_2, \ldots\}$ where $d(\psi_i, \psi_{i+1}) = \delta$ and $H_0 = (\psi_0 - \delta/2, \psi_0 + \delta/2]$ and compute the relative belief ratio

$$RB_\Psi((\psi_0 - \delta/2, \psi_0 + \delta/2] \,|\, x) = \frac{\Pi_\Psi((\psi_0 - \delta/2, \psi_0 + \delta/2] \,|\, x)}{\Pi_\Psi((\psi_0 - \delta/2, \psi_0 + \delta/2])}$$

to determine the evidence and the strength as

$$\begin{aligned} & Str_\Psi((\psi_0 - \delta/2, \psi_0 + \delta/2] \,|\, x) \\ = \;& \Pi_\Psi(RB_\Psi((\psi_i - \delta/2, \psi_i + \delta/2] \,|\, x) \leq \\ & RB_\Psi((\psi_0 - \delta/2, \psi_0 + \delta/2] \,|\, x) \,|\, x) \end{aligned}$$

- this "discretization" solves **many** problems with hypothesis assessment such as

$$\text{when } H_0 = (\psi_0 - \delta/2, \psi_0 + \delta/2] \text{ is true}$$
$$RB_\Psi((\psi_0 - \delta/2, \psi_0 + \delta/2] \,|\, x) \to 1/\Pi_\Psi((\psi_0 - \delta/2, \psi_0 + \delta/2]\dot{)} > 1$$
$$Str_\Psi((\psi_0 - \delta/2, \psi_0 + \delta/2] \,|\, x) \to 1$$
$$\text{when } H_0 = (\psi_0 - \delta/2, \psi_0 + \delta/2] \text{ is false}$$
$$RB_\Psi((\psi_0 - \delta/2, \psi_0 + \delta/2] \,|\, x) \to 0$$
$$Str_\Psi((\psi_0 - \delta/2, \psi_0 + \delta/2] \,|\, x) \to 0$$

- note that this discretization is not an ad hocracy, as it is an intrinsic part of the problem through δ, and a "good" application will have such a δ

- what do you do when you don't know δ? perform the analysis using several δ's to assess sensitivity

- with higher dimensional ψ such a discretization is still possible but it becomes computationally intractable as dimension rise so we need dimension reduction techniques

E

- suppose we want to estimate $\psi = \Psi(\theta)$

- the natural estimate is to choose the one which has the most evidence in its favor, namely,

$$\psi(x) = \arg \sup_{\psi \in \Psi(\Theta)} RB_\Psi(\psi \,|\, x)$$

and not this is the MLE from the inference base $(\{m_\psi : \psi \in \Psi(\Theta)\}, x)$

- so the MLE arises in a very natural way from a Bayesian context based on defining statistical evidence

- how about the accuracy of $\psi(x)$? for this quote the *plausible region*

$$\begin{aligned} Pl_\Psi(x) &= \{\psi : RB_\Psi(\psi \,|\, x) > 1\} \\ &= \text{ set of } \psi \text{ values with evidence in their favor} \end{aligned}$$

and its "size" (prior content) together with its posterior probability $\Pi_\Psi(Pl_\Psi(x) \,|\, x)$

- it is also possible to quote a γ-relative belief region

$$C_{\Psi,\gamma}(x) = \{\psi : RB_{\Psi}(\psi \,|\, x) \geq c_{\Psi,\gamma}(x)\}$$

where $c_{\Psi,\gamma}(x)$ is chosen to make $C_{\Psi,\gamma}(x)$ as small as possible while still satisfying $\Pi_{\Psi}(C_{\Psi,\gamma}(x)) \,|\, x) \geq \gamma$

- note - for a γ-relative belief region, we must have $\Pi_{\Psi}(Pl_{\Psi}(x) \,|\, x) \geq \gamma$ otherwise $C_{\Psi,\gamma}(x)$ will contain values of ψ for which evidence against has been found

- note - relative belief inferences respect the likelihood ordering induced by the inference base $(\{m_{\psi} : \psi \in \Psi(\Theta)\}, x)$