# Evans's PHYSTAT Discussion of $CL_s$

Michael Evans
Dept. of Statistical Sciences
University of Toronto
https://www.utstat.toronto.edu/mikevans/

May 8, 2024

# Background: what is the goal of a theory of statistics?

- a scientist wants answers to questions concerning a real-world object or quantity $\Psi$ (e.g. half-life of a neutron) and there are two statistical questions

**E**: what value does $\Psi$ take (estimation)?

**H**: does $\Psi = \psi_0$ (hypothesis assessment)?

- data is collected

- how should we reason, based on the data, to answer the question(s) of interest?

- central core concept: **statistical evidence**

- thesis: all statistical reasoning has to be clear about what this is and how it is to be used to determine the answers (inferences)

- evidential theories of statistics (Fisher, Jeffreys) versus decision-theoretic (Neyman, Wald, Lindley, Savage)

- the need to characterize statistical evidence has long been recognized without a generally accepted answer being presented

- why care? confidence in the reasoning process, replicability if the process can be shown to be sound, etc.

- A. Birnbaum's in the 60's and 70's was concerned with trying to characterize the concept in part via equivalence relations

- Birnbaum (1977) Synthese, offers something called the *confidence concept* in the context of comparing two hypotheses $H_0$ versus $H_1$

- the confidence concept is characterized by two *error probabilities*

$$\alpha = \text{probability of rejecting (accepting) } H_0 \ (H_1) \text{ when it is true (false)}$$
$$\beta = \text{probability of accepting (rejecting) } H_0 \ (H_1) \text{ when it is false (true)}$$

- then report $(\alpha, \beta)$ with the following interpretation

  rejecting $H_0$ constitutes strong evidence against $H_0$ (in favor of $H_1$) when $\alpha$ and $\beta$ are small

- if $\alpha(x), \beta(x)$ are p-values (associated with these error probabilities),

$$CL_s = \frac{\beta(x)}{1 - \alpha(x)}$$

and with a cut-off, e.g., $CL_s < 0.05$, obtain evidence against $H_1$

- be conservative in eliminating $H_1$, it can happen that just using $\beta(x)$ would result in finding evidence against $H_1$ which is not physical

- a reasonable approach to help avoid the problem but is it "correct"?

- this is similar to problems raised in: Mandelkern (2002) Setting confidence intervals for bounded parameters (with discussion). Stat. Sc. 17(2): 149-172.

**Example** *Poisson*$(\lambda)$ *with constraint* $\lambda \in (l_0, u_0)$ *where* $0 \le l_0 < u_0 \le \infty$

- $x_1, \ldots, x_n$ are *iid* Poisson$(\lambda)$ and it is known $\lambda \in (l_0, u_0)$ , how to form a confidence interval for $\lambda$ that satisfies this constraint without it being improper (absurd) with positive probability? ■

- addressing this issue: Evans, Liu, Moon, Sixta, Wei and Yang (2023) On some problems of Bayesian region construction with guaranteed coverages. Statistical Papers, doi.org/10.1007/s00362-023-01394-4.

- more generally address

> *Read (2000) "The goal of a search is to either exclude as strongly as possible the existence of a signal in its absence or confirm the existence of a true signal as strongly as possible while holding the probabilities of falsely excluding a true signal or falsely discovering a non-existent signal at or below specified level."*

# Measuring Statistical Evidence Using Relative Belief (Evans (2015))

**Ingredients:** data $x$, sampling model $\{f_\theta : \theta \in \Theta\}$, (proper) prior $\pi$, quantity of interest $\psi = \Psi(\theta)$ with prior $\pi_\Psi(\psi) = \sum_{\theta \in \Psi^{-1}\{\psi\}} \pi(\theta)$

- these ingredients are *falsifiable* via model checking and checking for prior-data conflict and there is no (unfalsifiable) loss function

**Principles (Axioms):**

(i) *Principle of Conditional Probability*: having observed $x$ prior beliefs are replaced by posterior beliefs

$$\pi_\Psi(\psi \,|\, x) = \sum_{\theta \in \Psi^{-1}\{\psi\}} \pi(\theta \,|\, x) = \sum_{\theta \in \Psi^{-1}\{\psi\}} \frac{\pi(\theta) f_\theta(x)}{m(x)}$$

where $m(x)$ = prior probability of observed data.

(ii) *Principle of Evidence*: If $\pi_\Psi(\psi \,|\, x) > (<) \pi_\Psi(\psi)$, then there is evidence in favor of (against) $\psi$ being the true value.

(iii) *Principle of Relative Belief*: base inferences on measuring evidence of $\psi$ being true by the relative belief ratio

$$RB_\Psi(\psi \,|\, x) = \frac{\pi_\Psi(\psi \,|\, x)}{\pi_\Psi(\psi)} \begin{cases} > 1 & \text{evidence in favor of } \psi \text{ being true,} \\ < 1 & \text{evidence against } \psi \text{ being true,} \\ = 1 & \text{no evidence either way.} \end{cases}$$

- some history, confirmation theory in the philosophy of science

> *Popper (1968) The Logic of Scientific Discovery, Appendix ix "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."*

- Berge Englert (https://phyweb.physics.nus.edu.sg/~phyebg/) a quantum physicist at NUS, BIT developed a similar approach

**H:** assess $H_0 : \Psi(\theta) = \psi_0$ via $RB_\Psi(\psi_0 \,|\, x)$

- how strong is the evidence?

> **There is no universal scale on which evidence is measured.**
> *Al-Labadi, Alzaatreh and Evans (2023) How to measure evidence and its strength: Bayes factors or relative belief ratios? arXiv:2301.08994*

- so $RB_\Psi(\psi_0 \,|\, x)$ needs to be calibrated and that is context dependent, basically use posterior probability to assess how strongly we believe what the evidence says

- $\psi_1$ not preferred to $\psi_2$ when $RB_\Psi(\psi_1 \,|\, x) \le RB_\Psi(\psi_2 \,|\, x)$ and

$$\Pi_\Psi(RB_\Psi(\psi \,|\, x) \le RB_\Psi(\psi_0 \,|\, x) \,|\, x)$$

is a measure of the strength of the evidence (not the evidence)

- don't have to assess the strength by one number

**E:** based on the ordering, estimate $\psi$ by

$$\psi(x) = \arg\sup RB_{\Psi}(\psi \mid x)$$

- error in estimate assessed by quoting the *plausible region*

$$Pl_{\Psi}(x) = \{\psi : RB_{\Psi}(\psi \mid x) > 1\} = \psi \text{ values with evidence in favor}$$

and measuring its "size" and posterior content

- $Pl_{\Psi}(x)$ only depends on Axioms (i) and (ii) so all *valid* estimates have the same accuracy.

- link with frequentism: $\psi(x) =$ MLE, $Pl_{\Psi}(x) =$ a likelihood region wrt the model $\{m(\cdot \mid \psi) : \psi \in \Psi(\Theta)\}$ (integrated out nuisance parameters)

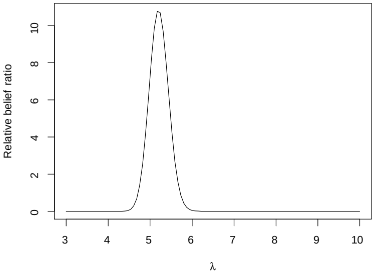
- $RB_{\Psi}$ is invariant under reparameterizations so all inferences are invariant and possess many other good (optimal) properties (see Evans (2015))

**Example** *Poisson*$(\lambda)$ *with constraint* $\lambda \in (l_0, u_0)$.

- gamma$_{rate}(\alpha_0, \beta_0)$ prior conditioned to $(l_0, u_0)$

- elicitation: interval $(l_1, u_1) \subset (l_0, u_0)$ is specified together with a probability $\gamma = \Pi((l_1, u_1))$ representing virtual certainty (e.g. $\gamma = 0.99$) this determines $(\alpha_0, \beta_0)$

- $(l_0, u_0) = (3, 10)$ and $(l_1, u_1) = (3.5, 9.5)$ with mode at $m_0 = (l_1 + u_1)/2$ and $\gamma = 0.99$ implies $\lambda \sim$ gamma$_{rate}(37.20, 5.57)$ conditioned to $(l_0, u_0)$



# Bias (Error probabilities)

- bias calculations are necessary as part of assessing the quality of a study

- e.g., should we accept the results of a statistical analysis that reported evidence against (in favor of) $H_0 : \Psi(\theta) = \psi_0$ if the prior probability of obtaining such evidence was $\approx 1$ when $H_0$ is true (false)?

**H**: **bias against** $H_0 : \Psi(\theta) = \psi_0$

$$M(RB_\Psi(\psi_0 \,|\, X) \leq 1 \,|\, \psi_0)$$

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

**H**: **bias in favor of** $H_0 : \Psi(\theta) = \psi_0$

$$\sup_{\psi : d_\Psi(\psi, \psi_0) \geq \delta} M(RB_\Psi(\psi_0 \,|\, X) \geq 1 \,|\, \psi)$$

the maximum prior probability of not getting evidence against $H_0$ when it is meaningfully false

**Example**: *Jeffreys-Lindley Paradox*

- $\bar{x} \sim N(\mu, \sigma_0^2/n)$ and $\mu \sim N(\mu_0, \tau_0^2)$

- assessing $H_0 : \mu = \mu_0$ then $RB(\mu_0 \,|\, \bar{x}) = BF(\mu_0 \,|\, \bar{x}) \to \infty$ as $\tau_0^2 \to \infty$

- could have classical p-value $2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_0|/\sigma_0)) \approx 0$ so contradiction between frequentism and Bayes

- $\Pi(RB(\mu \,|\, \bar{x}) \leq RB(\mu_0 \,|\, \bar{x}) \,|\, x) \to 2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_0|/\sigma_0))$ so evidence in favor is very weak in this situation (partial resolution)

- bias against $\to 0$ and bias in favor $\to 1$ as $\tau_0^2 \to \infty$

- general result: both biases converge to 0 as the amount of data increases and so bias can be controlled by design (not by choice of prior)

**E**: **bias against**

$$E_{\Pi_\Psi}\left(M(\psi \notin Pl_\Psi(X) \,|\, \psi)\right) = E_{\Pi_\Psi}(M(RB_\Psi(\psi \,|\, X) \leq 1 \,|\, \psi))$$

the prior probability that true value is not in the plausible region $Pl_\Psi(x)$

- $1 - E_{\Pi_\Psi}\left(M(\psi \notin Pl_\Psi(X) \,|\, \psi)\right)$ is the prior coverage prob. (Bayesian confidence) of $Pl_\Psi(x)$ wrt $m$

- typically there exist a $\psi_0 = \arg\sup M(RB_\Psi(\psi \,|\, X) \leq 1 \,|\, \psi)$ so

$$M(\psi \in Pl_\Psi(X) \,|\, \psi) \geq 1 - M(RB_\Psi(\psi_0 \,|\, X) \leq 1 \,|\, \psi_0)$$

gives a lower bound on the confidence of $Pl_\Psi(x)$ wrt $\{m(\cdot \,|\, \psi) : \psi \in \Psi\}$

**E**: **bias in favor**

$$E_{\Pi_\Psi}\left(\sup_{\psi' : d_\Psi(\psi', \psi) \geq \delta} M(\psi \notin Im_\Psi(X) \,|\, \psi')\right)$$

the prior probability that a meaningfully false value is not in the implausible region

- both biases $\to 0$ with increasing amounts of data

- these frequentist properties hold for **any** prior and depend only on **Axioms** (i) and (ii)

> **If bias assessments are held as being essential, then there are complementary roles for frequentism and Bayes.**

- general result Evans et al. (2023)

- put $C = \{x : RB_{\Psi}(\psi \,|\, x) = 1\} = \{x : m(x \,|\, \psi)$ is constant in $\psi\}$

**Theorem** *The plausible region for $\psi = \Psi(\theta)$ (i) never satisfies $Pl_{\Psi}(x) = \Psi(\Theta)$ and (ii) satisfies $Pl_{\Psi}(x) = \phi$ with prior probability 0 when $M(C) = 0$.*

- so, both unphysical outcomes and improper (absurd) confidence regions can be avoided by using appropriately conditioned priors

**Example** *Poisson*$(\lambda)$ *with constraint* $\lambda \in (l_0, u_0)$.

| $n$ | Bias against $H_0 : \lambda = 6.2$ with $\delta = 0.5$ |
|---|---|
| 1 | 0.287 |
| 10 | 0.193 |
| 20 | 0.085 |
| 50 | 0.045 |
| 100 | 0.001 |
| 500 | 0.000 |

Table: Bias against values for testing $H_0 : \lambda = 6.2$ for various sample sizes and meaningful differences.

- these are pure frequentist probabilities

| $n$ | Confidence level of $PI(\bar{x})$ using $\pi, \delta = 0.5$ (Bayes) |
|---|---|
| 1 | 0.581 (0.667) |
| 10 | 0.811 (0.840) |
| 20 | 0.843 (0.878) |
| 50 | 0.908 (0.935) |
| 100 | 0.950 (0.966) |
| 500 | 0.998 (0.999) |

Table: Frequentist (Bayesian) confidence that $PI(\bar{x})$ contains the true value in Example 4 for various sample sizes and meaningful differences.

- biases in favor require larger sample sizes to make small

# Conclusions and Question

- inferences are based on (dictated by) a clear definition of statistical evidence

- frequentism plays a role through the a priori control of error probabilities (the biases)

- unphysical results can be avoided

- can this help with the issue $CL_s$ is addressing in the context of particle physics in the sense that the argument has a sound foundational basis?