The Measurement of Statistical Evidence as the Basis for Statistical Reasoning

Michael Evans
University of Toronto
http://www.utstat.toronto.edu/mikevans/
paper: https://arxiv.org/abs/1906.09484

June 2019

The Basic Problem

- in a scientific context there are questions concerning an object of interest Ψ and data x has been collected believed to contain **evidence** concerning the answers

E **estimation** - provide an estimate $\psi(x)$ of Ψ together with an assessment of its accuracy based on the evidence

H **hypothesis assessment** - quote the evidence in favor of or against some specified value ψ_0 of Ψ *together with an assessment of the strength of the evidence*

How are we to reason from the data to answer **E** or **H**? based on characterizing/measuring statistical evidence

Evans (2015) Measuring Statistical Evidence Using Relative Belief

- Birnbaum (1964), Shafer (1976), Royall (1997), Thompson (2007), Aitkin (2010), Morey, Romeijn, and Rouder (2016), Vieland and Seok (2016),
- confirmation theory Salmon (1973), Achinstein (2001)

Valid Statistical Problems

- what problems are to be viewed as valid statistical problems?

Example - a sample x of size $n = 1$ is obtained with model $\{N(\mu, \sigma^2) : \mu \in R^1, \sigma^2 > 0\}$ and for $\Psi = (\mu, \sigma^2)$ a theory produces the absurd estimate $\psi(x) = (x, 0)$

- not a suitable problem to judge the correctness of a theory

- without constraints on what are core statistical problems, the provision of a suitable theory seems hopeless (?)

- this does not rule out compromises/hedges on noncore problems but these should be stated and motivated by a gold standard

- characteristics of core problems: data is collected via design + ...

The Role of Infinity

- many models incorporate infinities (small or large) but data is discrete (measured to finite accuracy and bounded)

- treating situations where infinities exist as the truth can cause anomalies

- here everything is essentially finite with infinity playing a role only through approximations expressed as limits

- Gauss (1831)

> *I protest against the use of infinite magnitude as something completed, which is never permissible in mathematics. Infinity is merely a way of speaking, the true meaning being a limit which certain ratios approach indefinitely close, while others are permitted to increase without restriction.*

Characteristics of a Theory of Statistical Reasoning

- in parallel with logical reasoning there are two aspects to a theory of statistical reasoning

(i) the ingredients (the premises)

(ii) the rules, or theory, of statistical inference (e.g. modus ponens) that are applied to the ingredients

- don't confound them (Aristotle - *valid* versus *sound* argument)

The Ingredients (Assumptions)

- what characteristics do we want these to have?

I_1 **Minimal** - the minimal ingredients needed to get a valid measure of evidence.

I_2 **Bias** - an assessment can be made to determine to what extent the chosen ingredients produce foregone conclusions to **E** or **H**.

I_3 **Falsifiable** - any (subjective) ingredient specified can be assessed against the (objective) data to see if it is contradicted.

$\vdots$

The Ingredients (Assumptions) Used Here

Model: $\{f_\theta : \theta \in \Theta\}$ a collection of conditional probability distributions for $x \in \mathcal{X}$ given θ such that the object of interest $\psi = \Psi(\theta)$ is specified by the true distribution that gave rise to x.

Prior: π a probability distribution on Θ.

Delta: δ the difference that matters so that $\text{dist}(\psi_1, \psi_2) \leq \delta$, for some distance measure dist, means that ψ_1 and ψ_2 are for practical purposes indistinguishable.

- reasonable to demand that the model and prior be elicited and that δ be provided (the hard stuff)

- this specifies a joint probability model for $\omega = (\theta, x) \sim \pi(\theta) f_\theta(x)$

Checking the Ingredients

- logically check for bias a priori

- to avoid bias an appropriate amount of data needs to be collected (discussed after the rules of inference)

- the model and the prior are *subjective* choices and need to be checked against the (*objective*) data x

- T is a minimal sufficient statistic for the model, then the joint factors as

$$\pi(\theta)f_\theta(x) = \pi(\theta \mid T(x))m_T(T(x))f(x \mid T(x))$$

- check the model first, based on $f(\cdot \mid T(x))$, then the prior, based on m_T

- **check the model**: is the observed data surprising for each f_θ?

- given the model is acceptable then

- **check the prior**: is the true value surprising for the prior?

- can you check a prior?

$$M_T(m_T(t) \leq m_T(T(x))) \rightarrow \Pi(\pi(\theta) \leq \pi(\theta_{true}))$$

as the amount of data increases, Evans and Jang (2011)

- small $M_T(m_T(t) \leq m_T(T(x)))$ indicates a problem with the prior

- prior-data conflict leads to robustness issues Al Labadi and Evans (2017)

- modifications and extensions Evans and Moshonov (2006), Nott et al (2018), etc.

- how to fix a bad prior? Evans and Jang (2011) one prior being weakly informative with respect to another and is data dependent only in the sense that conflict has been detected

The Rules of Statistical Inference

Evidence based - *answers to* **E** *and* **H** *are derived from a definition of how to characterize/measure evidence.*

- stated for a probability model $(\Omega, \mathcal{F}, P)$ when interest is in whether or not the event $A \in \mathcal{F}$ is true after observing $C \in \mathcal{F}$

$\mathbf{R}_1$: *Principle of conditional probability*: beliefs about A, as expressed initially by $P(A)$, are replaced by $P(A \mid C)$.

$\mathbf{R}_2$: *Principle of evidence*: the observation of C is evidence in favor of A when $P(A \mid C) > P(A)$, is evidence against A when $P(A \mid C) < P(A)$ and is evidence neither for nor against A when $P(A \mid C) = P(A)$.

$\mathbf{R}_3$: The evidence is measured quantitatively by the relative belief ratio

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

Keynes, Good, etc.

- $RB(A \mid C) > 1$ evidence for A, $RB(A \mid C) > 1$ evidence against A, $RB(A \mid C) = 1$ no evidence

- other *valid* (conforms to $\mathbf{R}_2$) measures exist, see Evans (2015)

- Bayes factor

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

- $RB(A \mid C) > 1$ iff $RB(A^c \mid C) < 1$

- for the statistical context: the relative belief ratio for $\psi = \Psi(\theta)$

$$RB_\Psi(\psi \mid x) = \lim_{\epsilon \to 0} \frac{\Pi(A_\epsilon \mid x)}{\Pi(A_\epsilon)} = \frac{\pi_\Psi(\psi \mid x)}{\pi_\Psi(\psi)}$$

where $\pi_\Psi(\cdot \mid x)$ is the posterior and π_Ψ is the prior

- when $\Pi_\Psi(\{\psi\}) = 0$ (continuous parameter) then could (should?) define

$$BF_\Psi(\psi \mid x) = \lim_{\epsilon \to 0} BF_\Psi(A_\epsilon \mid x) = RB_\Psi(\psi \mid x)$$

- differs from Jeffreys definition in the continuous case: modify prior to $\Pi_p = p\Pi_\psi + (1-p)\Pi$ where $p \in (0,1), \Pi_\psi(\Psi^{-1}\{\psi\}) = 1$ and usual Bayes factor works but ....

- a strictly increasing function of RB_Ψ gives the same inferences

- RB_Ψ is invariant under reparameterizations so all inferences are invariant

- Savage-Dickey $RB_\Psi(\psi \mid x) = \frac{m_\psi(x)}{m(x)}$ where m is the prior predictive and m_ψ is the conditional prior predictive given $\Psi(\theta) = \psi$

- so $RB_\Psi(\psi \mid x)$ is *proportional* to integrated likelihood (which doesn't measure evidence)

Inferences

E: ψ_1 not preferred to ψ_2 when $RB_\Psi(\psi_1 \mid x) \leq RB_\Psi(\psi_2 \mid x)$ so estimate

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

- error in estimate assessed via *plausible region*

$$Pl_\Psi(x) = \{\psi : RB_\Psi(\psi \mid x) > 1\}$$

- want $Pl_\Psi(x)$ "small" with high $\Pi_\Psi(Pl_\Psi(x) \mid x)$ for accuracy

- $Pl_\Psi(x)$ only depends on $\mathbf{R}_2$ so all *valid* estimates have the same accuracy

- could also quote a γ-credible region $C_{\Psi,\gamma}(x) = \{\psi : RB_\Psi(\psi \mid x) > k_\gamma\}$ provided $\gamma \leq \Pi_\Psi(Pl_\Psi(x) \mid x)$

Example *Prosecutor's Fallacy*

- a uniform probability distribution on a population of size N with trait left at a crime scene shared by $m < N$

- prosecutor cites rarity of trait and $P(\text{“has trait”} \mid \text{“guilty”}) = 1$ as evidence of guilt of a particular individual

- but $P(\text{“guilty”} \mid \text{"has trait”}) = 1/m$ could be small

- prosecutor is correct that possession of the trait is evidence of guilt

$$RB(\text{“guilty”} \mid \text{"has trait”})) = \frac{P(\text{“guilty”} \mid \text{“has trait”})}{P(\text{“guilty”})} = \frac{N}{m} > 1$$

- so $PI(\text{“has trait”}) = \{\text{“guilty”}\}$ but with posterior content $1/m$ so evidence weak when m is large

- MAP estimate is “not guilty” which doesn't reflect the evidence

Probabilities don't measure evidence!

- what about decisions?

H: assess $H_0 : \Psi(\theta) = \psi_0$ via $RB_\Psi(\psi_0 \mid x)$

- how strong is this evidence? compare $RB_\Psi(\psi_0 \mid x)$ to evidence for other ψ values to calibrate

- $\Pi_\Psi(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x)$ is a measure of the strength of the evidence

if $RB_\Psi(\psi_0 \mid x) < 1$ *and* $\Pi_\Psi(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x)$ *small, then strong evidence against* H_0

if $RB_\Psi(\psi_0 \mid x) > 1$ *and* $\Pi_\Psi(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x)$ *big, then strong evidence for* H_0

- as the amount of data increases:

when H_0 *false,* $RB_\Psi(\psi_0 \mid x) \to 0$ *and* $\Pi_\Psi(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x) \to 0$

when H_0 *true,* $RB_\Psi(\psi_0 \mid x) \to$ *value* >1 *and, when* ψ *is* **discrete**, $\Pi_\Psi(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x) \to 1$

- continuous case: H_0 is true whenever $\text{dist}(\psi_0, \psi_{true}) \leq \delta$ (E. G. Boring, Mathematical vs scientific significance, Psychol. Bull., I6, I9I9)

- this fixes the issue

- typically this makes the computations simpler

Jeffreys-Lindley Paradox and Bias

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

> *Would you 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* ≈ 1 *even when* H_0 *is true (false)?*

- bias calculations here only depend on the principle of evidence

Example - *Location-normal*
- $\bar{x} \sim N(\mu, \sigma_0^2/n)$ and $\mu \sim N(\mu_0, \tau_0^2)$ then $RB(\mu_0 \,|\, \bar{x}) \to \infty$ as $\tau_0^2 \to \infty$ (in this case $RB = BF$)

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

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

- need to discuss bias

- **note** $2(1-\Phi(\sqrt{n}|\bar{x}-\mu_0|/\sigma_0))$ doesn't satisfy $\mathbf{R}_2$ but

$$2(1-\Phi(\sqrt{n}|\bar{x}-\mu_0|/\sigma_0)) -$$
$$2(1-\Phi([\log(1+n\tau_0^2/\sigma_0^2)+(1+n\tau_0^2/\sigma_0^2)^{-1}(\bar{x}-\mu_0)^2/\tau_0^2]^{1/2})$$

does with cut-off 0

- rhs $\stackrel{a.s.}{\to} 0$ as $n\tau_0^2 \to \infty$ so the cut-off for $2(1-\Phi(\sqrt{n}|\bar{x}-\mu_0|/\sigma_0))$ to be evidence against also has to go to 0

- suppose $\sqrt{n}|\bar{x}-\mu_0|/\sigma_0 = 1.96$ and $\sigma_0 = 1, \mu_0 = 0, \tau_0^2 = 1$

n	p	evidence
10	$0.050 - 0.119 < 0$	against
50	$0.050 - 0.047 > 0$	in favor
100	$0.050 - 0.031 > 0$	in favor

- result is also true when the prior is Uniform$(-m, m)$ as $m \to \infty$ or as $n \to \infty$

- general resolution of Jeffreys-Lindley: measure and control bias

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

> **bias against** $M(RB_\Psi(\psi_0 \mid x) \leq 1 \mid \psi_0) =$ *prior probability of not getting evidence in favor of* H_0 *when it is true*
> **bias in favor** $\sup_{\psi : d(\psi, \psi_0) > \delta} M(RB_\Psi(\psi_0 \mid x) \geq 1 \mid \psi) =$ *maximum prior probability of not getting evidence against* H_0 *when it is false*

Example - *Location-normal*

- bias against $\longrightarrow 0$ and bias in favor $\longrightarrow 1$ as $\tau_0^2 \longrightarrow \infty$ (paradoxical?)

- in general, both biases converge to 0 as the amount of data increases and so bias can be controlled by design

Choose priors via elicitation, don't choose arbitrarily diffuse priors in an attempt to be "conservative", and design to avoid bias.

E: bias for estimating Ψ

bias against *the prior probability that true value is not in* $Pl_\Psi(x)$

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

- so $1 - E_{\Pi_\Psi}\left(M(\psi \notin Pl_\Psi(X) \,|\, \psi)\right)$ is the prior coverage probability (confidence) of $Pl_\Psi(x)$

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

- then $M(\psi \in Pl_\Psi(X) \,|\, \psi) \geq 1 - M(RB_\Psi(\psi_0 \,|\, X) \leq 1 \,|\, \psi_0)$ and a "pure" frequentist confidence when $\Psi(\theta) = \theta$ otherwise like a random effects model where random effects are the nuisance parameters

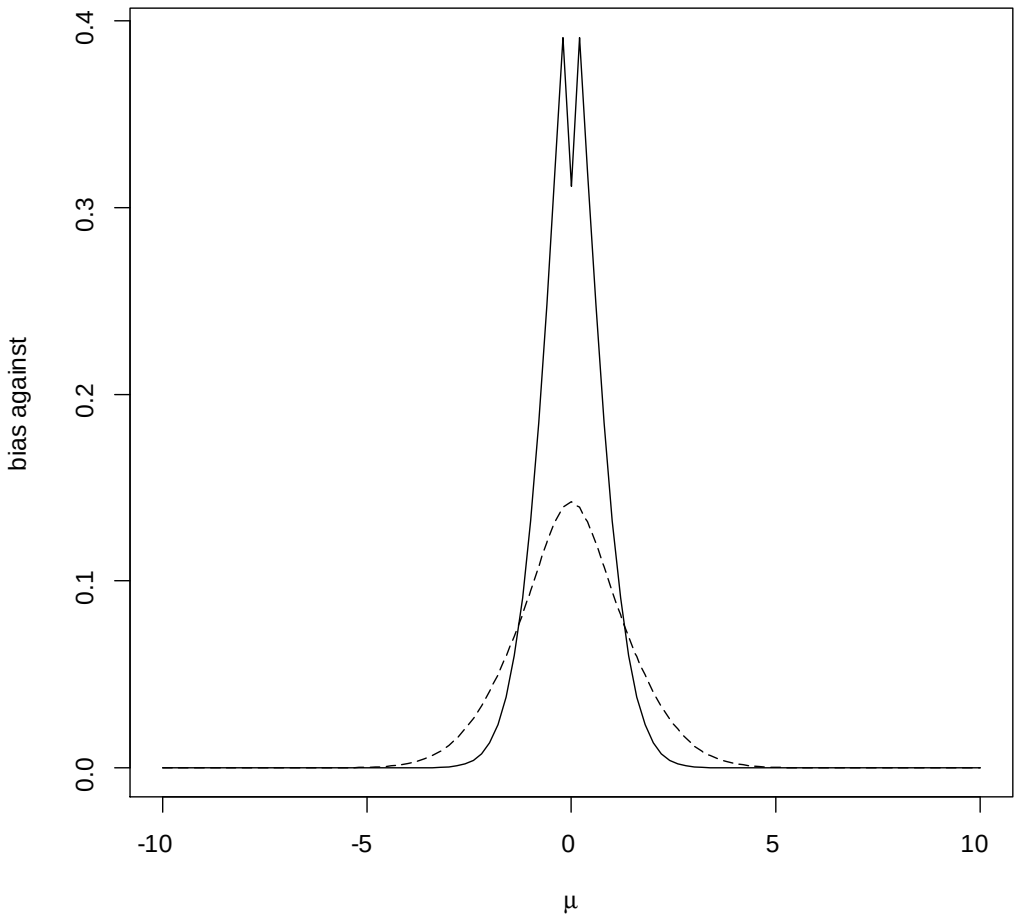


Figure: Plot of bias against $H_0 = \{\mu\}$ with a $N(0, 1)$ prior (- - -) and a $N(0, 0.01)$ prior (—) with $n = 5, \sigma_0 = 1$.

bias in favor *the prior probability that a meaningfully false value is not in the implausible region* $\mathrm{Im}_\Psi(x) = \{\psi : RB_\Psi(\psi_0 \mid x) < 1\}$

$$E_{\Pi_\Psi}\left(\sup_{\psi:d_\Psi(\psi,\psi_0)\geq\delta} M(\psi_0 \notin Im_\Psi(X) \mid \psi)\right)$$

$$= E_{\Pi_\Psi}\left(\sup_{\psi:d_\Psi(\psi,\psi_0)\geq\delta} M(RB_\Psi(\psi_0 \mid X) \geq 1 \mid \psi)\right)$$

- similar to the idea of measuring the accuracy of a confidence region via the probability of covering a false value

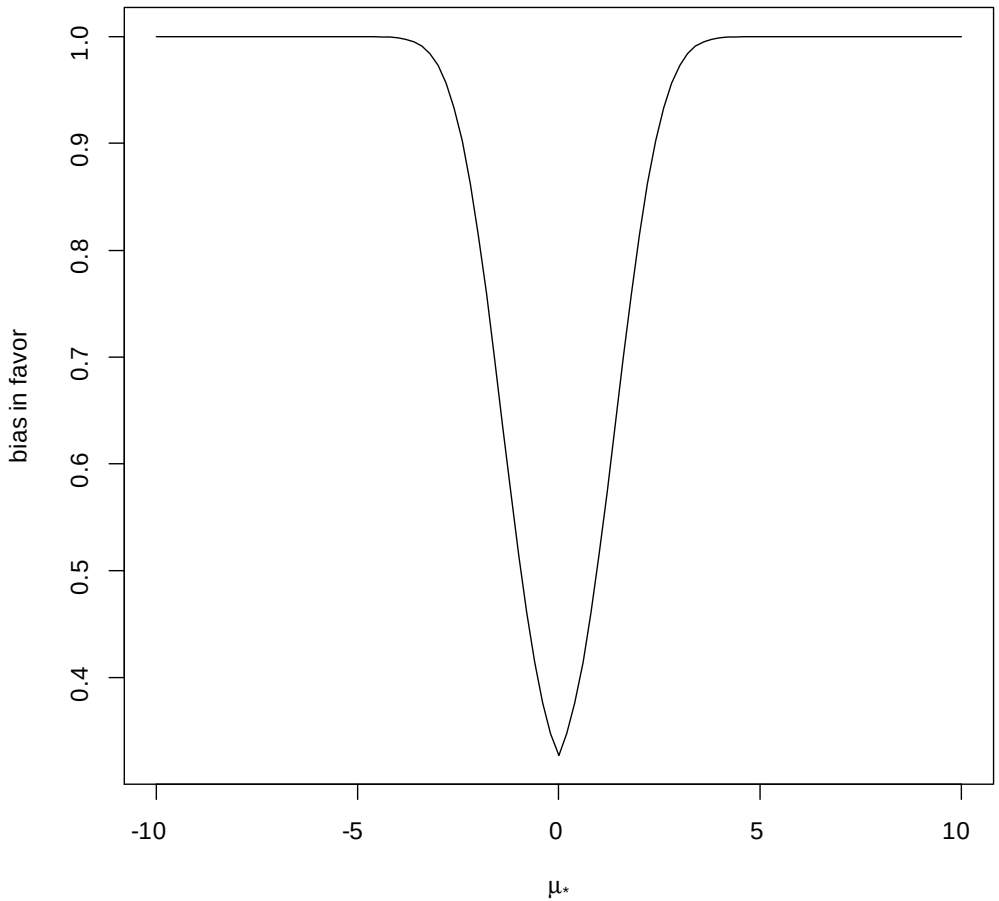


Figure: Bias in favor of μ maximized over $\mu \pm \delta$ based on a $N(0,1)$ prior with $\sigma_0 = 1, n = 20, \delta = 0.5$.

Some Properties

Are evidence measures incoherent?

- is it possible that $A \subset D$ but after observing C then $RB(A \mid C) > 1$ but $RB(D \mid C) < 1$? Yes!

- consider a population Ω and let C correspond those possessing a trait that exists only in A in subpopulation $A \cup B$

- $RB(A \mid C) = 1/P(C) > 1$ and $RB(B \mid C) = 0$ which implies $RB(A \cup B \mid C) = P(A \mid A \cup B)/P(C)$

- $RB(A \cup B \mid C) < 1$ iff $P(A \mid A \cup B) < P(C)$ or iff the probability of possessing the trait within the subpopulation is smaller than the probability of possessing the trait within the full population

Measuring evidence is different than measuring belief.

Theorem 1. Using the principle of evidence (i) the prior probability of getting evidence in favor of ψ_0 given that it is true, is greater than or equal to the prior probability of getting evidence in favor of ψ_0 given that ψ_0 is false and (ii) the prior probability of Pl_Ψ covering the true value is always greater than or equal to the prior probability of Pl_Ψ covering a false value.

- consider some other principle for establishing evidence and let

$$D(\psi) = \textit{data sets that don't lead to evidence in favor of } \psi$$
$$C(x) = \textit{which is the set of } \psi \textit{ values for which there is evidence in their favor after observing } x = \{\psi : x \notin D(\psi)\}$$

- for the principle of evidence these sets are respectively

$$R(\psi) = \{x : RB_\Psi(\psi \,|\, x) \leq 1\} \textit{ and } Pl(x) = \{\psi : x \notin R(\psi)\}$$

- alternative rules of interest satisfy

$$M(D(\psi) \,|\, \psi) \leq M(R(\psi) \,|\, \psi) \tag{1}$$

Theorem 2. If $\Pi_\Psi(\{\psi\}) = 0$, then $R(\psi)$ maximizes, among all rules satisfying (1), the prior probability of not obtaining evidence in favor of ψ when it is false and otherwise maximizes this probability among all rules satisfying $M(D(\psi) \mid \psi) = M(R(\psi) \mid \psi)$.

- if (1) holds for each ψ, then

$$E_{\Pi_\Psi}\left(M(\psi \in C(X)) \mid \psi)\right) \geq E_{\Pi_\Psi}\left(M(\psi \in Pl_\Psi(X)) \mid \psi)\right)$$

Theorem 3. If $\Pi_\Psi(\{\psi\}) = 0$ for all ψ, then Pl_Ψ maximizes, among all rules satisfying (1) for all ψ, the prior probability of not covering a false value and otherwise maximizes this probability among all C satisfying $M(\psi \notin C(X) \mid \psi) = M(\psi \notin Pl_\Psi(X) \mid \psi)$ for all ψ.

- results similar to Theorems 2 and 3 also exist when considering evidence in favor and are presented in Evans and Guo (2019)

- bias against and bias in favor for **H** are similar to frequentist size and power and for **E** are similar to frequentist confidence and accuracy

- in some cases these are the same

- if bias assessments are held as being essential, then there are complementary roles for frequentism and Bayes

> **Frequentism is concerned with assessing and controlling (via design) the quality of a study through the consideration of the possible data sets which might occur and their effects on inferences. Bayes is concerned with the inferences one draws based upon the actual data observed.**

Example: Prediction for Bernoulli Sampling

Diaconis and Skyrms (2018)

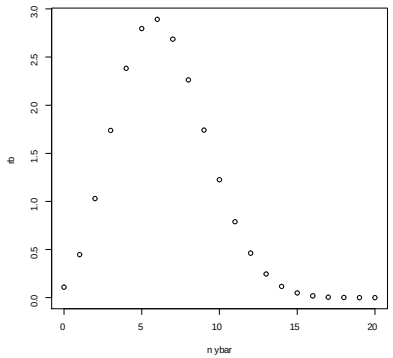
- $(x_1, \ldots, x_n) \sim \text{Bernoulli}(\theta), \theta \sim U(0,1)$

- MAP predictor of $(y_1, \ldots, y_f)$ maximizing posterior predictive $m_{n,f}((y_1, \ldots, y_f) \mid (x_1, \ldots, x_n))$ gives absurd answer

$$(y_1, \ldots, y_f) = \begin{cases} (0, \ldots, 0) & \text{if} \quad n\bar{x}/(n+f) \leq 1/2 \\ (1, \ldots, 1) & \text{if} \quad n\bar{x}/(n+f) \geq 1/2 \end{cases}$$

- note that when $f = n$, then $m_{n,n}((0, \ldots, 0) \mid (0, \ldots, 0)) \longrightarrow 1/2$ as $n \longrightarrow \infty$ and attempts to fix by modifying the prior (Jeffreys and Wrinch 1920's)

- relative belief predictor when $n = f = 20, n\bar{x} = 6$ is any sample with $f\bar{y} = 6$ and $Pl_n(x_1, \ldots, x_n) = \{(y_1, \ldots, y_f) : f\bar{y} = 2, 3, \ldots, 10\}$ has posterior content 0.893



- when $f = n, (x_1, \ldots, x_n) = (0, \ldots, 0)$ relative belief predictor is $(0, \ldots, 0), Pl_n(x_1, \ldots, x_n) \longrightarrow \{(0, 0, \ldots)\}$ and the posterior content of $Pl_n(0, \ldots, 0)$ converges to 1, Al-Labadi, Baskurt and Evans (2018)

Summary for Statistical Reasoning

1. Choose a model $\{f_\theta : \theta \in \Theta\}$.
2. Elicit a prior π.
3. Specify δ for characteristic of interest $\psi = \Psi(\theta)$.
4. Measure biases and determine appropriate amount of data x to collect.
5. Check the model against the data (modify if necessary).
6. Check the prior against the data (modify if necessary).
7. Inferences about ψ based on the principle of evidence.