Bayesian Epistemic Values: Focus on Surprise, Measure Probability

Probability-Possibility Transformations in Statistics

Julio Michael Stern*, **Carlos A.B. Pereira***
* Institute of Mathematics and Statistics
of the University of Sao Paulo
jstern@ime.usp.br

EBL-2011 COBAL-2011
MBR-2012 CBSF-2012
UniLog-2013

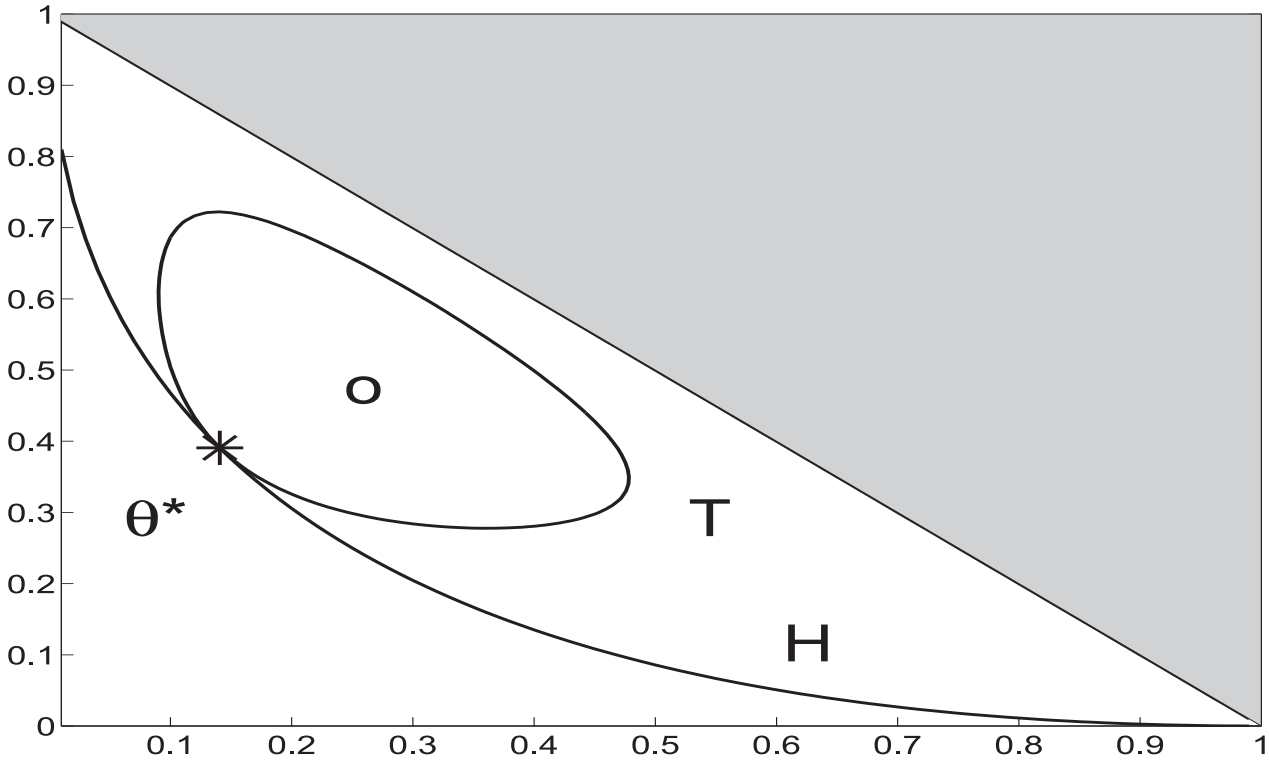
1- Previous Work of IME-USP Bayesian Group

Statistical significance, in empirical science, is the measure of belief or credibility or the truth value of an hypothesis.

1. Pereira and Stern (1999), Pereira et al. (2008): Statistical Theory of *e*-values - ev(H) or ev($H \mid X$) epistemic value of hypothesis H given de data X or evidence given by X in support of H.
2. Stern (2003, 2004), Borges and Stern (2007): "Logical" theory for composite *e*-valyes Compound Statistical Hypotheses in HDNF - Homogeneous Disjunctive Normal Form. (no such thing for *p*-values or Bayes factors)
3. Stern (2007a,b, 2008a,b, 2011a,b): Epistemological Framework given by Cognitive Constructivism.

Statistical Significance of Sharp Hypothesis H

States that the true value of the parameter, θ, of the sampling distribution, $p(x \mid \theta)$, lies in a low dimension set: The Hypothesis set, $\Theta_H = \{\theta \in \Theta \mid g(\theta) \leq 0 \wedge h(\theta) = 0\}$, has Zero volume (Lebesgue measure) in the parameter space.



Hardy-Weinberg Hypothesis

Bayesian setup

- $p(x \mid \theta)$: *Sampling distribution* of an observed (vector) random variable, $x \in \mathcal{X}$, indexed by the (vector) *parameter* $\theta \in \Theta$, regarded as a latent (unobserved) random variable.
- The model's joint distribution can be factorized either as the *likelihood function* of the parameter given the observation times the *prior* distribution on θ, or as the *posterior* density of the parameter times the observation's marginal density,

$$p(x, \theta) = p(x \mid \theta)p(\theta) = p(\theta \mid x)p(x) \ .$$

- $p_0(\theta)$: The *prior* represents our initial information.
- The *posterior* represents the available information about the parameter after 1 observation (unormalized potential),

$$p_1(\theta) \propto p(x \mid \theta)p_0(\theta) \ .$$

  Normalization constant $c_1 = \int_\theta p(x \mid \theta)p_0(\theta)d\theta$
- Bayesian learning is a recursive and comutative process.

- Hardy-Weinberg genetic equilibrium, see Pereira and Stern (1999).

  - n , sample size, x_1, x_3 , homozygote,
  - $x_2 = n - x_1 - x_3$, heterozygote count.
  - $\Theta = \{\theta \geq 0 \mid \theta_1 + \theta_2 + \theta_3 = 1\}$,
  - $H = \{\theta \in \Theta \mid \theta_3 = (1 - \sqrt{\theta_1}\,)^2\}$.

  $y = [0, 0, 0]$, Flat or uniform prior,
  $y = [-1/2, -1/2, -1/2]$, Invariant Jeffreys' prior,
  $y = [-1, -1, -1]$, Maximum Entropy prior.

$$p_0(\theta) \propto \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3} \ ;$$

Posterior density after observations $x = [x_1, x_2, x_3]$:

$$p_n(\theta \mid x) \propto \theta_1^{x_1+y_1} \theta_2^{x_2+y_2} \theta_3^{x_3+y_3} \ .$$

2- Full Bayesian Significance Test

- $r(\theta)$, the reference density, is a representation of no, minimal or vague information about the parameter θ. If $r \propto 1$ then $s(\theta) = p_n(\theta)$ and $\overline{T}$ is a HPDS.
- $r(\theta)$ defines the reference metric in Θ, $dl^2 = d\theta' J(\theta) d\theta$, directly from the Fisher Information Matrix,

$$J(\theta) \equiv -\mathrm{E}_{\mathcal{X}} \frac{\partial^2 \log p(x \mid \theta)}{\partial \theta^2} = \mathrm{E}_{\mathcal{X}} \left(\frac{\partial \log p(x \mid \theta)}{\partial \theta} \frac{\partial \log p(x \mid \theta)}{\partial \theta} \right).$$

- The *surprise function*, $s(\theta) = p_n(\theta)/r(\theta)$, measures changes in the posterior relative to the reference density.
- The 'hat' and 'star' superscripts indicate unconstrained and constrained (to the hypothesis H) maximal arguments and supremal surprise values, as follows:

$$\widehat{s} = \sup_{\theta \in \Theta} s(\theta) , \quad \widehat{\theta} = \arg\max_{\theta \in \Theta} s(\theta) ,$$
$$s^* = \sup_{\theta \in H} s(\theta) , \quad \theta^* = \arg\max_{\theta \in H} s(\theta) .$$

- The surprise function's (closed, upper-bound) *v-cut*, $T(v)$, its complement, the *highest surprise function set* (HSFS) above level v, $\overline{T}(v)$, and its *rim* (aka level-v set), $M(v)$, are

$$T(v) = \{\theta \in \Theta \,|\, s(\theta) \leq v\} \,, \quad \overline{T}(v) = \Theta - T(v) \,,$$

$$M(v) = \{\theta \in \Theta \,|\, s(\theta) = v\} \,.$$

- If the reference density the uniform (possibly improper) density, $r(\theta) \propto 1$, then $s(\theta) \propto p_n(\theta)$ and the HSFS are standard *highest probability density sets* (HPDS)
- The statistical model's *truth function*, $W(v)$, is the cumulative probability function up to surprise level v, $0 \leq v \leq \widehat{s}$. $\overline{W}(v)$ is its complement, $\overline{W}(v) = 1 - W(v)$, and $m(v)$ is its (generalized Schwartz) derivative,

$$W(v) = \int_{T(v)} p_n(\theta)\, d\theta \,, \quad m(v) = \frac{d}{dv} W(v) \,.$$

- Finally, the *e*-value for an hypothesis $H \subseteq \Theta$, $\text{ev}(H)$, aka the *epistemic value* of hypothesis H or the statistical evidence supporting H, and its complement, $\overline{\text{ev}}(H)$, are

$$\text{ev}(H) = W(v^*) \ , \quad \overline{\text{ev}}(H) = 1 - \text{ev}(H) \ .$$

  For the sake of simplicity, we use a relaxed notation for singleton arguments, that is, in the case of a *point hypothesis* $H = \{\theta^0\}$, writing $\text{ev}(\{\theta^0\}) = \text{ev}(\theta^0)$.
- The *e*-value of an hypothesis H is based on the most favorable case, $\text{ev}(H) = \text{ev}(\theta^*)$, a property that characterizes $\text{ev}(H)$ as a *possibilistic* ABC, (Abstract Belief Calculus).

3- Abstract Belief Calculus - ABC

- Darwiche, Ginsberg (1992).
- $\langle \Phi, \oplus, \oslash \rangle$, Support Structure;
  Φ , Support Function, for statements on $\mathcal{U}$;
  $\mathcal{U}$, Universe of valid statements;
  0 and **1**, Null and Full support values;
  $\oplus$, Support Summation operator;
  $\oslash$, Support Scaling or Conditionalization.

  $\otimes$, Support Unscaling, inverse of $\oslash$.
  $\langle \Phi, \oplus \rangle$, Partial Support Structure.

- $\oplus$, gives the support value of the disjunction of any two logically disjoint statements from their individual support values,

$$\neg(A \wedge B) \Rightarrow \Phi(A \vee B) = \Phi(A) \oplus \Phi(B) \ .$$

- $\oslash$, gives the conditional support value of B given A from the unconditional support values of A and the conjunction $C = A \wedge B$,

$$\Phi_A(B) = \Phi(A \wedge B) \oslash \Phi(A) \ .$$

- $\otimes$, unscaling: If Φ does not reject A,

$$\Phi(A \wedge B) = \Phi_A(B) \otimes \Phi(A) \ .$$

- Support structures for some belief calculi, Probability, Possibility, Classical Logic, Disbelief.
$a = \Phi(A)$, $b = \Phi(B)$, $c = \Phi(C = A \wedge B)$.

ABC	$\Phi(\mathcal{U})$	$a \oplus b$	$\mathbf{0}$	$\mathbf{1}$	$a \preceq b$	$c \oslash a$	$a \otimes b$
Pr	$[0,1]$	$a+b$	0	1	$a \leq b$	c/a	$a \times b$
Ps	$[0,1]$	$\max(a,b)$	0	1	$a \leq b$	c/a	$a \times b$
CL	$\{0,1\}$	$\max(a,b)$	0	1	$a \leq b$	$\min(c,a)$	$\min(a,b)$
DB	$\{0..\infty\}$	$\min(a,b)$	∞	0	$b \leq a$	$c-a$	$a+b$

- FBST setup: two belief calculi are in simultaneous use: ev constitutes a possibilistic (partial) support structure in the hypothesis space coexisting in harmony with the probabilistic support struct. given by the posterior probability measure in the parameter space.

4- Logic = Truth value of Composite Statements

- H in Homogeneous Disjunctive Normal Form;
  Independent statistical Models $j = 1, 2, \ldots$
  with stated Hypotheses $H^{(i,j)}$, $i = 1, 2 \ldots$
  Structures: $M^{(i,j)} = \{\Theta^j, H^{(i,j)}, p_0^j, p_n^j, r^j\}$.

$$\mathrm{ev}(H) = \mathrm{ev}\left(\bigvee_{i=1}^{q} \bigwedge_{j=1}^{k} H^{(i,j)}\right) =$$

$$\max_{i=1}^{q} \mathrm{ev}\left(\bigwedge_{j=1}^{k} H^{(i,j)}\right) =$$

$$W\left(\max_{i=1}^{q} \prod_{j=1}^{k} s^{*(i,j)}\right) ,$$

$$W = \bigotimes_{1 \le j \le k} W^j .$$

- Composition operators: max and $\otimes$ (Mellin convolution).
- Classical logic limit: If all $s^* = 0 \vee \widehat{s}$, ev $= 0 \vee 1$.

Wittgenstein's concept of Logic

- We analyze the relationship between the credibility, or truth value, of a complex hypothesis, H, and those of its elementary constituents, H^j, $j = 1 \ldots k$. This is the *Compositionality* question (ex. in analytical philosophy).
- According to Wittgenstein, (*Tractatus*, 2.0201, 5.0, 5.32):
  - Every complex statement can be analyzed from its elementary constituents.
  - Truth values of elementary statements are the results of those statements' truth-functions.
  - All truth-function are results of successive applications to elementary constituents of a finite number of truth-operations.
- Wahrheitsfunktionen, $W^j(s)$;
  Wahrheitsoperationen, $\bigotimes$, max.

Birnbaum's Logic for Reliability Eng.

- In reliability engineering, (Birnbaum, 1.4):
  "One of the main purposes of a mathematical theory of reliability is to develop means by which one can evaluate the reliability of a structure when the reliability of its components are known."
- Composition operations:
  - Series and parallel connections;
- Belief values and functions:
  - Survival probabilities and functions.
- There are **no** *logical rules* (composition operators) for the true values or functions used in classical statistics, *p*-values, or decision theoretic Bayesian statistics, Bayes factors.

5- Probability-Possibility Transformations

- Several important properties of $W(v)$ follow directly from the *nesting* property exhibited by the v-cuts that, in turn, give the integration range defining the truth function, see Dubois and Prade (1982),

$$u \leq v \Rightarrow T(u) \subseteq T(v) \Rightarrow W(u) \leq W(v) \; .$$

- Using this nesting property, it is easy to establish that $\text{ev}(H)$ has the desired properties of *consistency* with its underlying probability measure and *conformity* (to be similarly shaped) with its underlying surprise function, i.e.,

$$\text{Consistency:} \quad \text{ev}(H) \geq p_n(H) \; , \quad \forall \; H \subseteq \Theta \; ;$$

$$\text{Conformity:} \quad \text{ev}(\theta) \geq \text{ev}(\tau) \; \Leftrightarrow \; s(\theta) \geq s(\tau) \; , \quad \forall \; \theta, \tau \in \Theta \; .$$

A *plausibility measure*, $\mathrm{Pl}(H)$, is defined by its *basic probability assignment*, $m : 2^{\Theta} \mapsto [0,1]$, such that $\int_{S \subseteq \Theta} m(S) = 1$.
The *focal elements* of m are the subsets of the universe with non-zero basic pr.assignment, $\mathcal{F} = \{E \subseteq \Theta \,|\, m(E) > 0\}$.
Finally, the plausibility of $H \subseteq \Theta$, $\mathrm{Pl}(H)$, is defined as

$$\mathrm{Pl}(H) = \int_{E \in \mathcal{F} \,|\, E \cap H \neq \emptyset} m(E) \; .$$

Hence, $\mathrm{ev}(H)$ can be characterized as a plausibility function having v-cuts of the surprise function as focal elements, $\mathcal{F} = \{T(v), 0 \leq v \leq \widehat{s}\}$, while the basic probability density assigned to $T(v)$ is obtained integrating the posterior probability density over its rim, $m(v) = \int_{M(v)} p_n(\theta) d\theta$.

A plausibility function defines its dual *belief* function as

$$\mathrm{Bel}(\overline{H}) = \int_{E \in \mathcal{F} \,|\, E \subseteq \overline{H}} m(E) \;=\; 1 - \mathrm{Pl}(H) \; .$$

The *standard* possibility measure, $\pi(H)$, introduced by Dubois and Prade (1982, p.178), coincides (for the discrete case) with $ev(H)$ if $r(\theta) \propto 1$, the (trivial) uniform reference density.
- Distinct transformations were defined for the continuous case (should not have been an obstacle, but was a distraction).
- There are some traditional objections raised in decision theoretic Bayesian statistics against measures of statistical significance engendered by this transformation, namely,

(a) Lack of invariance.

(b) Not an ortodox decision theoretic procedure(?)
-An optimal point "represents" a composite hypothesis.

(c) No need for nuisance parameter elimination procedures.

(d) Epistemological interpretation of sharp hypotheses.

(e) Traditional understandings of significance tests as coverage (or not) of a point hypothesis, H', by a credibility interval of prescribed size. H' may be obtained by "pre-processing" (under permissible rules) the original statistical model.

(a) Invariance

ev(H) should not depend on the coordinate systems used to parameterize the statistical model.

- Reparameterization of H, i.e. of $h(\theta)$: Trivial.
- Consider a regular (bijective, integrable, a.s.cont.differentiable) reparameterization of Θ,

$$\omega = \phi(\theta) \quad , \quad \Omega_H = \phi(\Theta_H) \; .$$

The Jacobian of this coordinate transformation is

$$J(\omega) = \left[\frac{\partial\,\theta}{\partial\,\omega}\right] = \left[\frac{\partial\,\phi^{-1}(\omega)}{\partial\,\omega}\right] = \begin{bmatrix} \frac{\partial\,\theta_1}{\partial\,\omega_1} & \cdots & \frac{\partial\,\theta_1}{\partial\,\omega_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial\,\theta_n}{\partial\,\omega_1} & \cdots & \frac{\partial\,\theta_n}{\partial\,\omega_n} \end{bmatrix} ,$$

and the surprise funcion in the new coordinates is

$$\widetilde{s}(\omega) = \frac{\widetilde{p}_n(\omega)}{\widetilde{r}(\omega)} = \frac{p_n(\phi^{-1}(\omega))\,|J(\omega)|}{r(\phi^{-1}(\omega))\,|J(\omega)|} \; .$$

$$\text{Hence,} \quad \widetilde{s}^* = \sup_{\omega \in \Omega_H} \widetilde{s}(\omega) = \sup_{\theta \in \Theta_H} s(\theta) = s^* \ ,$$

$$T(s^*) \mapsto \phi(T(s^*)) = \widetilde{T}(\widetilde{s}^*) \ , \text{ and}$$

$$\widetilde{\text{Ev}}(H) = \int_{\widetilde{T}(\widetilde{s}^*)} \widetilde{p}_n(\omega) d\omega = \int_{T(s^*)} p_n(\theta) d\theta = \text{ev}(H) \ , \quad \text{Q.E.D.}$$

Box and Tiao (1965, p.1470): *"It seems that we cannot hope for invariance for a genuine measure of credibility. It needs to be remembered that invariance under transformations and virtues are not synonymous. For problems which should not be invariant under transformation, a search for invariance serves only to guarantee inappropriate solutions."*

- This claim went undisputed in the statistical literature!!
- Possibilistic mesures *"must be thought of as a very informal way of testing."* Harrison (1997, Sec.8.6.7, p.256,257).

(b) Decision Theoretic Analysis

M.R.Madruga, L.G.Esteves, S.Wechsler (2001):

- Loss function, $\Lambda : \{Accept, Reject\} \times \Theta \mapsto \mathcal{R}$,

  $\Lambda(R, \theta) = a\,\mathbf{1}(\theta \in T)$, $\Lambda(A, \theta) = b + c\,\mathbf{1}(\theta \in \overline{T})$.
- Minimum loss: Accept H iff $\text{ev}(H) > \varphi = \frac{b+c}{a+c}$.

ev(H) leads to an ortodox decision theoretic procedure, even if a single point, the constrained optimal estimator $\theta^* = \arg\max_H s(\theta)$, "represents" the entire hypothesis set!!

Traditionally, Bayesian procedures use only integral operators, never a maximization operator. Notice that a classical p-value, as the e-value, is defined using both operations.
- p-values have pseudo-possibilitic characteristics.

(c) No need for Nuisance Parameter Elimination

- Dimensionality reduction technique.
  - Allows the "reduction" of H to dimension zero (e).
  - Difficult problems can be solved with simple devices, like the Pickett N-525-T Statistics Slide Rule.
- The FBST does not follow the nuisance parameters elimination paradigm, working in the original parameter space, in its full dimension, breaking away from both the frequentist and the decision theoretic Bayesian tradition.
- NPE? - That's not a bug, that's a feature!
  How does a (theoretical) bug become a feature?

  Raymond Chen (Microsoft): *"One thing you quickly learn in application compatibility is that a bug once shipped gains the status of a feature, because you can be pretty sure that some program somewhere relies on it."*
- The FBST allways requires the use of numerical optimization and integration methods (MC, MCMC, etc.)

Epistemological interpretation of sharp hypotheses

- In decision theoretic Bayesian statistics,
  Bayes Factors are related to "betting odds" for H.
  However, a sharp hypothesis has zero probability!
  That is the ZPP - The Zero Probability Paradox.
- The ZPP creates several technical dificulties, like Lindley's paradox, and motivates many ad-hoc fixes, like artficial or special purpose priors (caveat emptor).
- Sharp hypotheses make no sense in the decision theoretic epistemological (de Finettian) framework.
- Sharp hipotheses are fully supported in the Cognitive Constructivism + FBST epistemological framework.
- Deeply entangled with question (e).
  Interesting back-propagation to Possibility th. literature!

Bibliography

- W.Borges, J.M.Stern (2007). The Rules of Logic Composition for the Bayesian Epistemic e-Values. *Logic J. IGPL*, 15, 401-420.
- M.Diniz, C.A.B.Pereira, J.M.Stern (2011). Unit Roots: Bayesian Significance Test. *Communications in Statistics - Theory and Methods,* 40, 23, 4200-4213 .
- D.Dudois H.Prader (1982). On Several Representations of an Uncertain Body of Evidence. p. 167-181 in M.Gupta, E.Sanchez, *Fuzzy Information and Decision Processes,* North-Holland.
- M.Lauretto, C.A.B.Pereira, J.M.Stern, S.Zacks (2003). Full Bayesian Signicance Test Applied to Multivariate Normal Structure Models. *Brazilian J.of Prob.& Statistics,* 17, 147-168.
- M.R.Madruga, L.G.Esteves, S.Wechsler (2001). On the Bayesianity of Pereira-Stern Tests. *Test*, 10, 291–299.
- C.A.B.Pereira, J.M.Stern (1999). Evidence and Credibility: Full Bayesian Significance Test Precise Hypotheses. *Entropy*, 1, 69-80.
- C.A.B.Pereira, S.Wechsler, J.M.Stern (2008). Can a Significance Test be Genuinely Bayesian? *Bayesian Analysis,* 3, 79-100.

- J.M.Stern (2004). Paraconsistent Sensitivity Analysis for Bayesian Significance Tests. SBIA'04, *LNAI*, 3171, 134–143.
- J.M.Stern (2007a). Cognitive Constructivism, Eigen-Solutions, and Sharp Statistical Hypotheses. *Cybernetics and Human Knowing*, 14, 9-36.
- J.M.Stern (2007b). Language and the Self-Reference Paradox. *Cybernetics and Human Knowing*, 14, 71-92.
- J.M.Stern (2008a). Decoupling, Sparsity, Randomization, and Objective Bayesian Inference. *C& HK*, 15, 49-68.
- J.M.Stern (2008b). *Cognitive Constructivism and the Epistemic Significance of Sharp Statistical Hypotheses.* Tutorial book for MaxEnt 2008, The 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. July 6-11 of 2008, Boracéia, São Paulo, Brazil.
- J.M.Stern (2011a). Constructive Verification, Empirical Induction, and Falibilist Deduction: A Threefold Contrast. *Information,* 2011, 2, 635-650.
- J.M.Stern (2011b). Symmetry, Invariance and Ontology in Physics and Statistics. *Symmetry,* 2011, 3, 611-635.