The Bayesian Approach to Discovery

Jim Berger

Duke University

PHYSTAT2011
CERN
January 17, 2011

Outline

- A simple example of Bayesian testing
- A Bayesian formulation of the generic HEP problem
- Multiplicity
  - Of hypotheses
  - Of cuts
- Choosing priors for Bayesian testing
- A random thought about unfolding

A Simple Example of Bayesian Testing

Data: $N = \#$ events observed in time T that are characteristic of Higgs boson production in LHC particle collisions.

Statistical Model: N has density

$$\text{Poisson}(N \mid s + b) = \frac{(s+b)^N e^{-(s+b)}}{N!}\,,$$

where

- s is the mean rate of production of Higgs events in time T;
- b is the (known) mean rate of production of events with the same characteristics from background sources in time T.

To test: $H_0 : s = 0$ vs $H_1 : s > 0$. (H_0 corresponds to 'no Higgs.')

P-value: $P(N \geq N_{\text{observed}} \mid b, s = 0) = \sum_{j=N}^{\infty} \text{Poisson}(j \mid 0 + b)$

Case 1: $p = 0.00025$ if $N = 7$, $b = 1.2$

Case 2: $p = 0.025$ if $N = 6$, $b = 2.2$.

- Those who understand p-values know their use is difficult:
  Luc Demortier: In any search for new physics, a small p value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better explanation of the observations than the null hypothesis can one make a convincing case for discovery.
- Bayesian analysis directly measures if the alternative hypothesis provides a better explanation.

Sequential testing: This is actually a sequential experiment, so p should be adjusted to account for multiple looks at the data. Bayesian analysis does not need such a correction.

Bayes factor of H_0 to H_1: *ratio of likelihood under* H_0 *to average likelihood under* H_1 (or "odds" of H_0 to H_1)

$$B_{01}(N) = \frac{\text{Poisson}(N \mid 0+b)}{\int_0^\infty \text{Poisson}(N \mid s+b)\pi(s)\, ds} = \frac{b^N\ e^{-b}}{\int_0^\infty (s+b)^N\ e^{-(s+b)}\pi(s)\, ds}\,.$$

Subjective approach: Choose $\pi(s)$ subjectively (e.g., using the standard physics model predictions of the mass of the Higgs).

Objective approach: Choose $\pi(s)$ to be the 'intrinsic prior' (not discussed here) $\pi^I(s) = b(s+b)^{-2}$. (Note that this prior is proper and has median b.)

Bayes factor: is then given by

$$B_{01} = \frac{b^N\ e^{-b}}{\int_0^\infty (s+b)^N\ e^{-(s+b)} b(s+b)^{-2}\ ds} = \frac{b^{(N-1)}\ e^{-b}}{\Gamma(N-1, b)}\,,$$

where Γ is the incomplete gamma function.

Case 1: $B_{01} = 0.0075$ (recall $p = 0.00025$)

Case 2: $B_{01} = 0.26$ (recall $p = 0.025$)

Posterior probability of the null hypothesis: The objective choice of prior probabilities of the hypotheses is $\Pr(H_0) = \Pr(H_1) = 0.5$, in which case

$$\Pr(H_0 \mid N) = (1 + B_{01}^{-1})^{-1} .$$

Case 1: $\Pr(H_0 \mid N) = 0.0075$ (recall $p = 0.00025$)
Case 2: $\Pr(H_0 \mid N) = 0.21$ (recall $p = 0.025$)

Complete posterior distribution: is given by

- $\Pr(H_0 \mid N)$, the posterior probability of null hypothesis
- $\pi(s \mid N, H_1)$, the posterior distribution of s under H_1

A useful summary of the complete posterior is $\Pr(H_0 \mid N)$ and C, a (say) 95% posterior credible set for s under H_1.

Case 1: $\Pr(H_0 \mid N) = 0.0075$; $C = (1.0, 10.5)$

Case 2: $\Pr(H_0 \mid N) = 0.21$; $C = (0.2, 8.2)$

Note: For testing precise hypotheses, confidence intervals alone are *not* a satisfactory inferential summary. Note also that Bayes gives a joint summary, so no correction of the confidence interval is needed after testing.

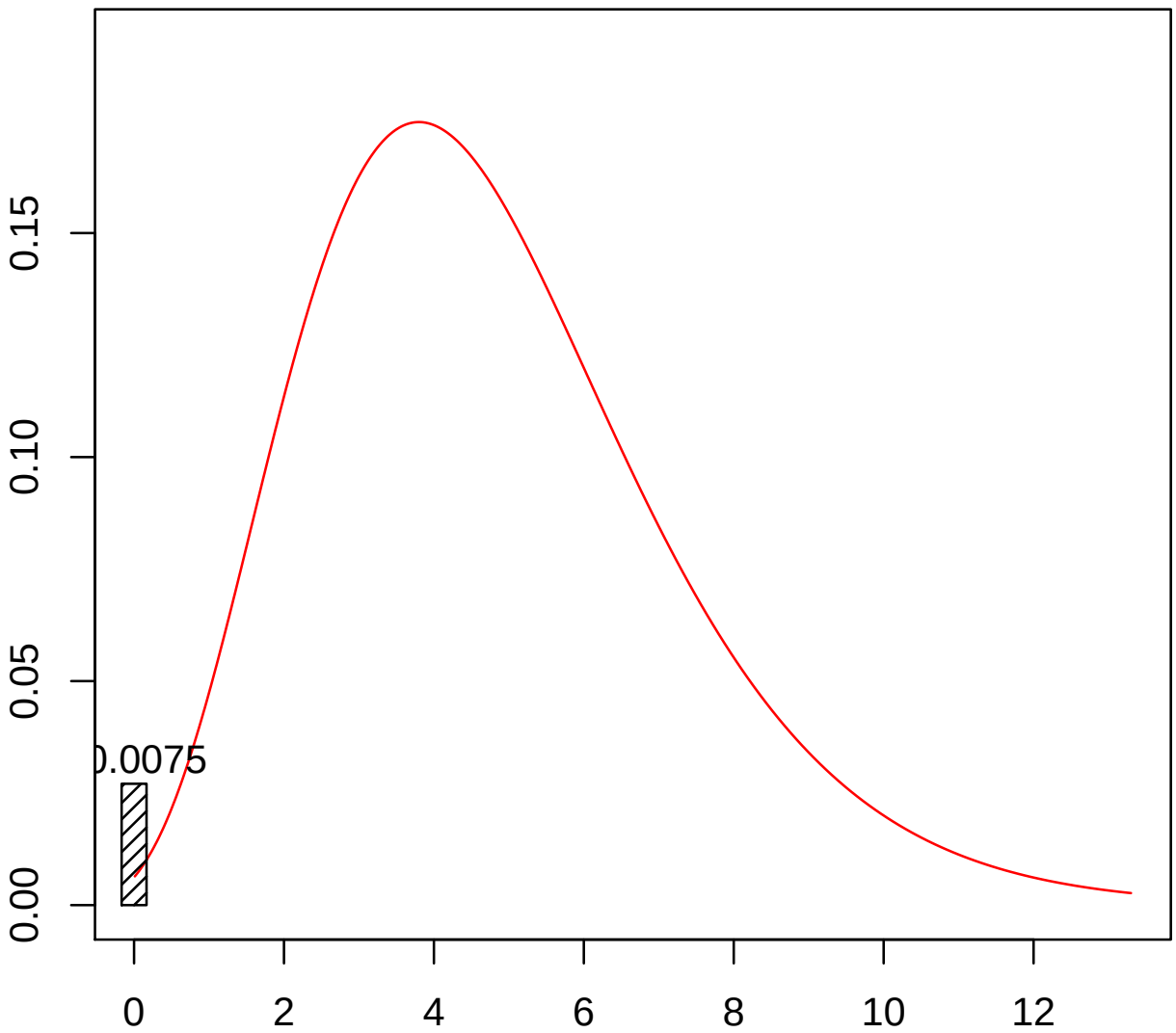

Figure 1: $\Pr(H_0 \mid N)$ (the vertical bar), and the posterior density for s given N and H_1.

Is the discrepancy between p-values and Bayes factors due to choice of the prior?

A lower bound on the likelihood ratio (or Bayes factor): choose $\pi(s)$ to be a point mass at $\hat{s}$, yielding

$$B_{01}(N) = \frac{\text{Poisson}(N \mid 0+b)}{\int_0^\infty \text{Poisson}(N \mid s+b)\pi(s)\, ds} \geq \frac{\text{Poisson}(N \mid 0+b)}{\text{Poisson}(N \mid \hat{s}+b)}$$

$$= \min\{1, \left(\frac{b}{N}\right)^N e^{N-b}\}\,.$$

Case 1: $B_{01} \geq 0.0014$ (recall $p = 0.00025$)

Case 2: $B_{01} \geq 0.11$ (recall $p = 0.025$)

Note: This use of *robust Bayesian analysis* was done in Edwards, Lindman and Savage (1963) and Berger and Sellke (1987); many generalizations followed; indeed, there is now the Society of Imprecise Probabilities.

A Bayesian Formulation of the Basic HEP Problem

The statistical model (following Richard Lockhart's Banff II writeup):

- N is the observed Poisson number of events.
- The events are independent and each has characteristics ('marks' in the Poisson process world) X_i, $i = 1, \ldots, N$.
- Under H_0: *background only*,
  - the mean of N is b,
  - the density of the X_i is $f_b(x) > 0$.
- There may be a signal Poisson process with mean s and density $f_s(x)$.
- Under H_1: *background + signal*,
  - the mean of N is $b + s$,
  - the density of the X_i is $(\gamma f_b(x) + (1-\gamma) f_s(x))$, where $\gamma = \frac{b}{(b+s)}$.
- Consider the case where $f_b(x)$ and $f_s(x)$ are known but b and s are unknown.

Bayes factor of H_1 to H_0 for priors $\pi_0(b)$ and $\pi_1(b,s) = \pi_0(b)\pi_1(s \mid b)$:

$$B_{10} = \frac{\int_0^\infty \int_0^\infty (b+s)^N e^{-(b+s)} \prod_{i=1}^N [\gamma f_b(x_i) + (1-\gamma) f_s(x_i)] \pi_1(b,s)\, dsdb}{\int_0^\infty b^N e^{-b} \prod_{i=1}^N [f_b(x_i)]\, \pi_0(b) db}$$

$$= \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^N \left[1 + \frac{s f_s(x_i)}{b f_b(x_i)}\right] \pi_0(b)\pi_1(s \mid b)\, dsdb}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db} .$$

Note that, if b is known, this becomes

$$B_{10} = \int_0^\infty e^{-s} \prod_{i=1}^N \left[1 + \frac{s f_s(x_i)}{b f_b(x_i)}\right] \pi_1(s \mid b)\, ds .$$

Priors: Intrinsic priors are $\pi_0^I(b) = b^{-1/2}$ (note that it is improper) and $\pi_1^I(s \mid b) = b(s+b)^{-2}$ (note that it is proper).

Note: Ignoring the densities f_s and f_b and basing the answer solely on N is equivalent to assuming that $f_s \equiv f_b$.

Look-elsewhere Concerns are Automatically Handled

I. Multiple Hypotheses: Suppose N_j of the X_i are in bin B_j, $j = 1, \ldots, M$, and that we assume we have only densities $f_s(B_j)$ and $f_b(B_j)$. Then

$$B_{10} = \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^M \left[1 + \frac{s f_s(B_j)}{b f_b(B_j)}\right]^{N_j} \pi_0(b)\pi_1(s \mid b)\, dsdb}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db}\,.$$

Suppose $f_s(B_j)$ gives probability one to some unknown bin B (the signal could occur in only one bin), with each bin being equally likely. Then

$$\begin{aligned} B_{10} &= \frac{E^B\left[\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^M \left[1 + \frac{s f_s(B)}{b f_b(B_j)}\right]^{N_j} \pi_0(b)\pi_1(s \mid b)\, dsdb\right]}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db} \\ &= \frac{1}{M}\sum_{j=1}^M \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \left[1 + \frac{s}{b f_b(B_j)}\right]^{N_j} \pi_0(b)\pi_1(s \mid b)\, dsdb}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db}\,, \end{aligned}$$

so that the results from each H_j:*signal in* B_j are downweighted by $1/M$.

II. Multiple Cuts:

- Bayesian analysis has no easy way to adjust for cuts.
  - Cuts produce subsets of the data, but there is no Bayesian way to analyze separate subsets and combine them (unless the subsets are independent).
  - If each cut produces an N_j, one could legitimately consider the joint distribution of the N_j, but that is too hard.
  - One could, of course, consider the union of the data from the cuts.
- Are cuts necessary if one does a Bayesian analysis?

Suppose $f_s(x) = 0$ for $x \in \Omega^c$. We could cut on Ω, but do we need to do so?

$$B_{10} = \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{\{i:x_i\in\Omega\}} \left[1 + \frac{s f_s(x_i)}{b f_b(x_i)}\right] \pi_0(b)\pi_1(s \mid b)\, ds db}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db}$$

$$= \int_0^\infty B_{10}(x_\Omega \mid b)\pi_0(b \mid x_\Omega, x_{\Omega^c}) db \quad \text{(follows from algebra, not probability)},$$

where x_Ω (x_{Ω^c}) is the data in Ω (in Ω^c). Bayes indeed uses the cut data to find the Bayes factor given b, but uses all the data to learn more about b.

An example of the difficulty in frequentist control of multiplicity:

One tests $H_{0i} : \mu_i = 0$ versus $H_{1i} : \mu_i > 0$, $i = 1, ..., m$.

Data: X_i are normally distributed with mean θ_i, variance 1, and correlation ρ.

If $\rho = 0$, one can just do individual tests at level α/m (Bonferroni) to obtain an overall error probability of α.

If $\rho > 0$, harder work is needed:

- Choose an overall decision rule, e.g., "declare μ_i to be the signal if X_i is the largest value and $X_i > K$."
- Compute the corresponding error probability:

$$\alpha = E^Z \left[1 - \Phi \left(\frac{K - \sqrt{\rho} Z}{\sqrt{1-\rho}} \right)^m \right],$$

  where Φ is the standard normal cdf and Z is standard normal.

Note that this gives (essentially) the Bonferroni correction when $\rho = 0$, and converges to $1 - \Phi[K]$ as $\rho \to 1$.

Choosing priors for "common parameters" in testing

If evidence-based priors are not available, here are some guidelines:

- Gold standard: if there are parameters in each hypothesis that have the same group invariance structure, one can use the right-Haar priors for those parameters (even though improper) (Berger, Pericchi and Varshavsky, 1998)
- Silver standard: if there are parameters in each hypothesis that have the same scientific meaning (e.g. background mean), reasonable default priors (e.g. reference priors or the constant prior 1) can typically be used.
- Bronze standard: to try to obtain parameters that have the same scientific meaning (beware the "fallacy of Greek letters"), one strategy often employed is to orthogonalize the parameters, i.e., reparameterize so that the partial Fisher information for those parameters is zero.

Choosing priors for non-common parameters

If evidence-based priors are not available, here are some guidelines:

- Vague proper priors are horrible (related to the Jeffreys-Lindley paradox): for instance, if $X \sim N(\mu, 1)$ and we test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ with a Uniform$(-c, c)$ prior for θ, the Bayes factor is

$$B_{01}(c) = \frac{f(x \mid 0)}{\int_{-c}^{c} f(x \mid \mu)(2c)^{-1} d\mu} \approx \frac{2c\, f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu) d\mu}$$

  for large c, which depends dramatically on the choice of c.

- Improper priors are problematical, because they are unnormalized; is

$$B_{01} = \frac{f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu)(1) d\mu} \quad \text{or} \quad B_{01} = \frac{f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu)(2) d\mu} \, ?$$

- Robust solution: if one can specify a plausible range $c_1 \leq c \leq c_2$, look at $B_{01}(c)$ over this range and hope the conclusion is robust. (Not obvious for higher dimensional parameters, but there is a literature.)

Case 1: $\pi(\mu)$ is Uniform$(0, 10)$ (e.g., known upper limit on μ)

- Observe $x = 2$: $p = 0.025$, while $\Pr(H_0 \mid x = 2) = 0.54$
- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 1.3 \times 10^{-3}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 6.0 \times 10^{-8}$

Case 2: $\pi(\mu)$ is Normal$(4, 1)$ (arising from a previous experiment)

- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 4.7 \times 10^{-4}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 5.8 \times 10^{-8}$

Case 3: $\pi(\mu)$ is a point mass at 4 (the prediction of a new theory).

- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 3.4 \times 10^{-4}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 1.1 \times 10^{-7}$

Conservative conversion of p to Pr$(H_0 \mid x)$*:* $\Pr(H_0 \mid x) = (1 + (-ep \log p)^{-1})^{-1}$:

- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 8.8 \times 10^{-4}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 5.7 \times 10^{-8}$

Various proposed default priors for non-common parameters

- Fractional priors (O'Hagan): use a fraction γ of the model likelihood (usually $\gamma =$'parameter dimension'/'sample size') as the prior, with $L(\theta)^{1-\gamma}$ as the likelihood.
- Intrinsic priors (Berger, Pericchi and others): generate priors from "training samples" (either actual subsets of the data, or imaginary data generated under the null model).
- Conventional priors that have at least some nice properties: e.g., Zellner-Siow priors for linear models are
  - invariant to scale changes in covariates
  - consistent (the true model will be selected as $n \to \infty$)
  - information-consistent (e.g., will reject as t or F statistics $\to \infty$)
  - coherent (roughly, are logically connected)
- Various efforts at 'predictive matching' priors.
- Approximations (such as BIC); these can capture part of the prior influence, but not all.

Comment on Unfolding

Shyamalkumar (unpublished) had the following interesting result about finding $\pi(\theta)$ such that

$$m_\pi(x) = \int f(x \mid \theta)\pi(\theta)\, d\theta$$

is as close as possible to an estimated $\hat{m}(x)$:

- choose any initial $\pi_0(\theta)$ that has support everywhere;
- iteratively compute

$$\pi_l(\theta) = \int \pi_{l-1}(\theta \mid x)\hat{m}(x)dx\,, \quad \text{where } \pi_{l-1}(\theta \mid x) = \frac{\pi_{l-1}(\theta)f(x \mid \theta)}{\int \pi_{l-1}(\theta)f(x \mid \theta)d\theta}\,.$$

Fact: $\pi^*(\theta) = \lim_{l\to\infty} \pi_l(\theta)$ is the density for which $m_\pi(x)$ is as close as possible to $\hat{m}(x)$ in Kullback-Leibler divergence.

A Possibly Interesting Implementation via Particle Filtering:

- Represent $\pi_l(\theta)$ by a collection of *particles* $\{\theta_i\}$ with weights $\{w_i^{(l)}\}$. (Initialize with a random sample $\{\theta_i\}$ from $\pi_0(\theta)$, so the initial weights are equal.)
- Then $\pi_l(\theta \mid x)$ would be the same collection of particles but with modified weights

$$w_i^{(l)}(x) = \frac{w_i^{(l-1)} f(x \mid \theta_i)}{\sum_j w_j^{(l-1)} f(x \mid \theta_j)} ,$$

  and $\pi_l(\theta)$ would be the same collection of particles but with weights

$$w_i^{(l)} = \int w_i^{(l)}(x) \hat{m}(x) dx \, .$$

- As one progresses one will need to add new particles adapting to the evolving density, but there are likely techniques in particle filtering for doing this.

Miscellaneous Comments

- Bayesian combining of independent evidence is straightforward: for instance
$$B_{10}(x_1, x_2 \mid \pi) = B_{10}(x_2 \mid \pi(\cdot \mid x_1))B_{10}(x_1 \mid \pi)\,.$$
- Concerning the Bayesian approach to the look-elsewhere effect:
  - Not all look-elsewhere effects (e.g., observation of sequential data) require Bayesian adjustment.
  - Bayesian correction can always be formulated as choice of prior probabilities of models.

Thanks!