

A Bayes Factor Example: Poisson Discovery
 Joel Heinrich—University of Pennsylvania
 February 3, 2009

1 Introduction

This note is an introduction to the Bayes factor via an example based on the following problem: the main experiment observes events with a Poisson rate that derives from a signal of cross section s (with acceptance ϵ) and background b . Information about ϵ and b derives from Poisson subsidiary measurements:

$$\begin{aligned} n &\sim \text{Pois}(\epsilon s + b) && \text{(main measurement)} \\ y &\sim \text{Pois}(tb) && \text{(subsidiary background measurement)} \\ z &\sim \text{Pois}(u\epsilon) && \text{(subsidiary acceptance measurement)} \end{aligned}$$

Constants t and u relate the number of events observed in the subsidiary measurements to the expectation for background and acceptance in the main experiment. The ϵ parameter, not constrained to be ≤ 1 , is actually acceptance times integrated luminosity. See [1] for a review of the performance of parameter estimation methods (upper limits on the parameter of interest s) for this problem.

The question addressed here is one of significance, or hypothesis testing: having observed (n, y, z) , one would like to assess the evidence for discovery. The traditional frequentist threshold for discovery in HEP is 5σ significance. The Bayesian approach to hypothesis testing involves calculating the Bayes factor. Further discussion of Bayes factors can be found in [2].

As the reader will discover, prior selection for the Bayes factor in the triple Poisson example treated here is quite challenging. It is not anticipated that the Bayes factor will be much used at CDF, which is approaching the end of data taking, but there is serious discussion at the LHC of using the Bayes factor in addition to the more familiar p-value approach. The rationale is that providing both a frequentist and a Bayesian answer instills confidence in cases where they agree, and caution when they disagree. As the conclusions in this note indicate, the Bayes factor may prove too problematical for widespread use except in the simplest of cases.

2 The Likelihood Ratio

The frequentist approach (i.e. significance) requires the selection of a statistic that quantifies the disagreement between the observed data and the null

hypothesis, which is taken in this case to be $s = 0$. The likelihood ratio λ is the preferred statistic in the frequentist approach, in this case it is

$$\lambda = \frac{\text{likelihood with } s = 0, \text{ maximized w.r.t. } \epsilon, b}{\text{likelihood maximized w.r.t. } s, \epsilon, b}$$

and therefore $0 \leq \lambda \leq 1$. The likelihood L is given by

$$L(s, \epsilon, b) = \frac{e^{-(\epsilon s + b)} (\epsilon s + b)^n}{n!} \frac{e^{-tb} (tb)^y}{y!} \frac{e^{-u\epsilon} (u\epsilon)^z}{z!}$$

The maximizations required to calculate the likelihood ratio can be easily carried out analytically in this simple example. The result is

$$\lambda = \frac{L(0, z/u, (n+y)/(t+1))}{L(u(n-y/t)/z, z/u, y/t)}$$

where it notable that $\epsilon = z/u$ is obtained both under the restricted maximization (with $s = 0$) and the unrestricted maximization.

We then have

$$\lambda = \frac{\left(\frac{n+y}{t+1}\right)^{n+y}}{n^n \left(\frac{y}{t}\right)^y}$$

We have not restricted the range of maximization for the denominator of the likelihood ratio to $s \geq 0$; should that restriction be desired, we have $\lambda = 1$ when $n \leq y/t$.

Thus, in this example, the likelihood ratio is independent of z (the result of the subsidiary measurement for the acceptance). Under frequentist repetitions of (n, y, z) for the null hypothesis, the probability distribution of neither n nor y depend on the true value of ϵ , so the distribution of λ under the null will only depend on the assumed true value of b . We would prefer the distribution of λ to also be independent of b , and it turns out that this is asymptotically true; for $b \gg 1$, $-2 \ln \lambda$ is approximately χ^2 distributed, here with one degree of freedom, independent of the actual value of b .

3 Bayes Factor

The Bayes factor B in this case is by definition

$$B = \frac{\text{likelihood with } s = 0, \text{ marginalized over } \epsilon, b}{\text{likelihood marginalized over } s, \epsilon, b}$$

which is quite similar in spirit to the definition of the likelihood ratio, with Bayesian marginalization replacing maximization. Here marginalization implies multiplying the likelihood by a prior, then integrating.

Before selecting priors, we will reparametrize s and b in terms of new parameters ρ and μ where $s = \mu\rho/\epsilon$ and $b = \mu(1 - \rho)/(t + 1)$. The inverse transformation is given by $\mu = \epsilon s + (t + 1)b$ and $\rho = \epsilon s / (\epsilon s + (t + 1)b)$, so $0 \leq \rho \leq 1$ and $\mu \geq 0$.

This reparametrization is a typical approach, taking advantage of the duality between two independent Poissons (variables n and y in our problem) and a single Poisson on the overall $n + y$ with a binomial determining the distribution of n for fixed $n + y$. Jeffreys uses this approach in his example ‘‘Test for consistency of two Poisson parameters’’ worked out in [3], which is similar to our problem.

The Bayes factor, with this reparametrization, is then

$$B = \frac{\text{likelihood with } \rho = 0, \text{ marginalized over } \epsilon, \mu}{\text{likelihood marginalized over } \rho, \epsilon, \mu}$$

The likelihood as a function of the new parameters is

$$L(\rho, \epsilon, \mu) = \frac{e^{-\mu \frac{t\rho+1}{t+1}} [\mu \frac{t\rho+1}{t+1}]^n}{n!} \frac{e^{-\mu t \frac{1-\rho}{t+1}} [\mu t \frac{1-\rho}{t+1}]^y}{y!} \frac{e^{-u\epsilon} (u\epsilon)^z}{z!}$$

The marginalization step requires selection of priors. We assume that the joint prior factors in the form $\pi(\rho, \mu, \epsilon) = \pi(\rho)\pi(\mu, \epsilon)$, and observe that the marginalization integrals over μ and ϵ cancel between numerator and denominator of the Bayes factor (for any choice of $\pi(\mu, \epsilon)$ whatsoever) yielding

$$\frac{1}{B} = \int_0^1 (t\rho + 1)^n (1 - \rho)^y \pi(\rho) d\rho$$

3.1 an attractive but ultimately unsuccessful choice of prior

A convenient form (see discussion in sec 3.2) for the remaining prior is

$$\pi(\rho) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \rho^{\alpha-1} (1 - \rho)^{\beta-1}$$

which is a beta distribution, a proper distribution defined on the interval $[0, 1]$ having mean $\frac{\alpha}{\alpha + \beta}$ and variance $\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$. This yields ([4] eq 15.3.1)

$$\frac{1}{B} = \frac{\Gamma(\alpha + \beta)\Gamma(y + \beta)}{\Gamma(\beta)\Gamma(y + \alpha + \beta)} F(-n, \alpha; y + \alpha + \beta; -t)$$

where F is the hypergeometric function. A prior flat in ρ is the special case with $\alpha = \beta = 1$.

In sec 3.5 the beta prior will be shown to be problematical when t is large, leading us to pick a more general form for the prior in sec 3.6.

3.2 prior discussion

We would like to see what our choice of prior for ρ implies for the prior for s and b . As B was independent of priors for ϵ and μ , we did not need to select any prior for those parameters. We will now try the choice

$$\pi(\rho, \mu, \epsilon) \propto \rho^{\alpha-1}(1-\rho)^{\beta-1}\mu^{\gamma-1}\pi(\epsilon)$$

and calculate the transformed prior. Calculating the Jacobian yields

$$d\mu d\rho = \frac{\epsilon(t+1)}{\epsilon s + (t+1)b} ds db$$

and we obtain

$$\pi(s, b, \epsilon) \propto \frac{s^{\alpha-1}b^{\beta-1}\epsilon^\alpha\pi(\epsilon)}{[\epsilon s + (t+1)b]^{\alpha+\beta-\gamma}}$$

By picking $\gamma = \alpha + \beta$, we can reduce it to

$$\pi(s, b, \epsilon) \propto s^{\alpha-1}b^{\beta-1}\epsilon^\alpha\pi(\epsilon)$$

which is a familiar form, often used in the estimation of the parameter of interest s . Choosing the beta distribution as the prior for ρ was motivated by this connection.

However, this is puzzling, as it seems that via a simple transformation we have succeeded in using an improper prior for s , a parameter that is marginalized under only one of the hypotheses. But under the alternative ($s = 0$) hypothesis, we have $\mu = (t+1)b$, and the prior $\pi(\mu, \epsilon) \propto \mu^{\alpha+\beta-1}\pi(\epsilon)$ transforms to $\pi(b, \epsilon) \propto b^{\alpha+\beta-1}\pi(\epsilon)$. This is inconsistent with the form we obtained for $\pi(s, b, \epsilon)$. So the transformation to variables (s, b, ϵ) is not a success; we end up with different priors for b under the two hypotheses. At best, then, we have provided some plausibility for choosing a beta distribution as the prior for ρ ; it can't be derived rigorously from our favored (improper) prior for s .

3.3 the hypergeometric function

For complex w with $|w| < 1$, the hypergeometric function (specifically Gauss' hypergeometric function) is defined by its power series

$$F(a, b; c; w) = 1 + \frac{abw}{c} \frac{1}{1!} + \frac{a(a+1)b(b+1)}{c(c+1)} \frac{w^2}{2!} + \frac{a(a+1)(a+2)b(b+1)(b+2)}{c(c+1)(c+2)} \frac{w^3}{3!} + \dots$$

and outside that range by analytic continuation. We don't have to worry about the case where c is a nonpositive integer in our example. For the

special case of nonpositive integer a , the series terminates, and F becomes a polynomial in w of degree $-a$ (and the restriction $|w| < 1$ is lifted).

The closely related function

$${}_2F_0(a, b; ; w) = 1 + ab \frac{w}{1!} + a(a+1)b(b+1) \frac{w^2}{2!} + a(a+1)(a+2)b(b+1)(b+2) \frac{w^3}{3!} + \dots$$

only converges when a or b is a nonpositive integer, otherwise it diverges for all $|w| > 0$.

3.4 Bayes factor vs p -value

The Bayes factor gives the ratio of posterior to prior odds in the Bayesian approach, and is intended to be interpreted directly, rather than be converted into a p -value. Nevertheless, it is interesting to see what Bayes factor corresponds to $\sim 5\sigma$ p -value in this problem. There are, however, many methods of p -value calculation that can be applied to this problem. The conditioning method [5] is one possibility. In this method, one calculates the p -value conditioned on the observed $n + y$, which converts the probability into that of a binomial distribution. One obtains

$$p_{\text{cond}} = \frac{1}{(1+t)^{n+y}} \sum_{j=0}^y \frac{(n+y)!}{(n+y-j)! j!} t^j$$

for the p -value. The conditioning method is perhaps not optimum for discovery purposes (being more conservative than necessary due to the discrete nature of the binomial distribution), but it leads to a simple analytic expression for the p -value. We also try the supremum method [6] using the (restricted) likelihood ratio statistic, which has more power, but is more difficult to implement. (Both the conditioning method and the supremum method are fully frequentist.)

The value of B corresponding to 5σ discovery in this example is about 10^{-5} . This is illustrated with a few examples in the following table for the case with $t = 4$, $\alpha = 1$, $\beta = 1$:

n	y	B	p_{cond}	p_{sup}
10	1	0.87e-5	9.22e-7	2.42e-7
11	2	1.15e-5	10.7e-7	5.95e-7
14	5	1.00e-5	6.80e-7	5.37e-7
18	10	0.86e-5	4.24e-7	2.42e-7
23	18	1.03e-5	3.74e-7	2.42e-7
26	23	0.92e-5	2.92e-7	1.60e-7
28	27	1.11e-5	3.22e-7	1.76e-7
7	12	0.733	0.068	0.045

Examining more extensive tables, one does not find any strong disagreements between the Bayesian approach and the frequentist in this example; nothing stands out that would lead one to draw different conclusions from the two approaches. The last line in the table illustrates the fact that a 5% p -value provides very little evidence against the null hypothesis, but in HEP this is already well understood.

3.5 the large t limit

There is one sign of trouble though; the behavior of our expression for B as $t \rightarrow \infty$. The likelihood ratio has a well defined limit when we let $y = b_0 t$, $b_0 > 0$ representing the true value of b

$$\lim_{t \rightarrow \infty} \lambda = e^{n-b_0} \left(\frac{b_0}{n} \right)^n$$

which is exactly the value we would have obtained with a fixed background b_0 . Similarly, the limiting value of p_{cond} is given by

$$\lim_{t \rightarrow \infty} p_{\text{cond}} = e^{-b_0} \sum_{j=n}^{\infty} \frac{b_0^j}{j!}$$

as expected. The limit of our Bayes factor, however, is

$$\lim_{t \rightarrow \infty} B = \lim_{t \rightarrow \infty} \frac{\Gamma(\beta)(b_0 t + \beta)^\alpha}{\Gamma(\alpha + \beta) {}_2F_0(-n, \alpha; ; -1/b_0)} = \infty$$

which blows up because ([4] eq 6.1.47)

$$\frac{\Gamma(y + \alpha + \beta)}{\Gamma(y + \beta)} \simeq (y + \beta)^\alpha$$

as y becomes large (it becomes an equality when $\alpha = 1$).

Thus, as t becomes large, the null ($s = 0$) hypothesis becomes more and more favored. This, of course, is pathological for our intended use, but it can be explained. After reparametrization, our null hypothesis, which had been $s = 0$, has been changed to $\rho = 0$, and because of our definition of ρ as the fraction of *all* events in the main and subsidiary background channels due to the signal; letting $t \rightarrow \infty$ automatically forces $\rho \rightarrow 0$, even if s is finite. A flat prior, for example, for ρ is therefore inconsistent with the large t limit (independent of the true value of s). The Bayes factor responds rationally, but its logic was perhaps not initially obvious.

Interestingly, neither λ nor p_{cond} suffer from $s = 0$ vs $\rho = 0$ ambiguity in this example. The likelihood ratio is invariant under reparametrization, and

although p_{cond} is conditional on $n + y$, which becomes infinite in the limit $t \rightarrow \infty$, the desired limiting behavior for p_{cond} is still obtained.

Because of the remaining freedom in choosing the prior for ρ , we can attempt to include the extra insight that our prior should be concentrated near $\rho = 0$ for large t . This can be done in our beta distribution ρ -prior, for example, by choosing $\beta = \beta_1 t + \beta_0$ for constant β_1 and β_0 . This modifies the limit of B to

$$\lim_{t \rightarrow \infty} B = \frac{(b_0/\beta_1 + 1)^\alpha}{{}_2F_0(-n, \alpha; ; -1/(b_0 + \beta_1))}$$

which seems more promising. But there is still trouble when $b_0 \ll \beta_1$ and t is large, since in that case the Bayes factor loses its dependence on b_0 —it seems that a simple beta prior for ρ is not adequate to cover all cases. We need a different prior.

3.6 a modified prior

Reference [7] computes a Bayes factor for a simplified version of this problem with $\epsilon = 1$ fixed and b also fixed (i.e. single Poisson); the prior for s is constructed using an algorithm that yields an “intrinsic” prior $\pi(s) \propto (s + b)^{-2}$, where the label “intrinsic” identifies the particular procedure employed. The intrinsic prior is not unique, being itself derived from an underlying estimation prior which is not unique.

The failure of our beta prior for ρ in the calculation of the Bayes factor is not actually unexpected or surprising. Quoting from [7]:

In Bayesian hypothesis testing and model selection, however, determination of suitable prior distributions is considerably more challenging, in part because it is typically the case that improper prior distributions cannot be used (or at least have to be used very carefully). Use of ‘vague proper priors’ (another staple of many Bayesians in estimation problems) is even worse, and will typically give nonsensical answers in testing and model selection. There has thus been a huge effort in statistics to derive objective (or at least conventional) priors for use in hypothesis testing and model selection.

In our triple Poisson case, the $s + b$ of the intrinsic prior of [7] generalizes to $\epsilon s + b$, which is proportional to $t\rho + 1$ in our transformed (ρ, μ) system. This suggests generalizing our prior for ρ to

$$\pi(\rho) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)F(-\delta, \alpha; \alpha + \beta; -t)} \rho^{\alpha-1} (1 - \rho)^{\beta-1} (t\rho + 1)^\delta$$

which is proper for any value of δ provided $\alpha > 0$ and $\beta > 0$. When δ is negative this has the behavior we desire, forcing the prior to peak at zero when t is large. With this modified prior, the Bayes factor for our case becomes

$$B = \frac{\Gamma(\beta)\Gamma(y + \alpha + \beta)F(-\delta, \alpha; \alpha + \beta; -t)}{\Gamma(\alpha + \beta)\Gamma(y + \beta)F(-n - \delta, \alpha; y + \alpha + \beta; -t)}$$

For judicious choices of α , β , and δ , we will see that this expression for B has reasonable behavior in the limit of large t . For simplicity, we will begin with a few special cases where δ is an integer. The case $\delta = 0$ has already shown to be pathological, so we start with $\delta = -1$, $\alpha = \beta = 1/2$, which yields

$$B = \frac{\sqrt{\pi}\Gamma(y + 1)(1 + t)^{-1/2}}{\Gamma(y + 1/2)F(-n + 1, 1/2; y + 1; -t)}$$

whose limiting value at large t is given by

$$\lim_{t \rightarrow \infty} B = \frac{\sqrt{\pi b_0}}{{}_2F_0(-n + 1, 1/2; ; -1/b_0)}$$

which seems fine (we ignore the case $n = 0$ for now). Selecting $\delta = -2$, $\alpha = \beta = 1$, we get

$$B = \frac{(y + 1)(1 + t)^{-1}}{F(-n + 2, 1; y + 2; -t)}$$

whose limiting value at large t is given by

$$\lim_{t \rightarrow \infty} B = \frac{b_0}{{}_2F_0(-n + 2, 1; ; -1/b_0)}$$

which agrees exactly with the result of [7].

A case that behaves badly is $\delta = -1$, $\alpha = \beta = 1$. The resulting Bayes factor is

$$B = \frac{(y + 1) \ln(1 + t)}{t F(-n + 1, 1; y + 2; -t)}$$

which diverges logarithmically at large t .

So our modified prior is more promising than the beta prior, although care still needs to be taken. The special cases examined above used integer values of δ ; for non integer values, one can apply transformations ([4] eqs 15.3.4–5) to the hypergeometric functions in the numerator and denominator that yields

$$B = \frac{\Gamma(\beta)\Gamma(y + \alpha + \beta)F(\alpha + \beta + \delta, \alpha; \alpha + \beta; t/(t + 1))}{\Gamma(\alpha + \beta)\Gamma(y + \beta)(1 + t)^{n + \alpha + \delta}F(-n - \delta, y + \beta; y + \alpha + \beta; t/(t + 1))}$$

which is a suitable form when $0 < -\alpha - \delta < n$ and t is large to speed numeric convergence of the series representation of the hypergeometric function. However, when n is large, this power series will be numerically unstable, as successive terms become large in magnitude but opposite in sign. For $0 \leq n < -\alpha - \delta$

$$B = \frac{\Gamma(\beta)\Gamma(y + \alpha + \beta)F(\alpha + \beta + \delta, \alpha; \alpha + \beta; t/(t + 1))}{\Gamma(\alpha + \beta)\Gamma(y + \beta)F(y + n + \alpha + \beta + \delta, \alpha; y + \alpha + \beta; t/(t + 1))}$$

converges faster at large t , and may be the only usable form when n is large. The limiting value is given by

$$\lim_{t \rightarrow \infty} B = \frac{\Gamma(-\alpha - \delta)}{\Gamma(-\delta)U(\alpha, 1 + n + \alpha + \delta, b_0)}$$

This may be derived by applying the transformation given in eq 15.3.6 of [4] and observing that the limit as $t \rightarrow \infty$ yields the confluent hypergeometric function U of eq 13.1.3 in [4]. The behavior at large t is divergent when $\alpha + \delta \geq 0$. Equations 13.5.6–12 in [4] give the behavior of $U(a, b, z)$ for small $|z|$. For $n = 0$, we have

$$\lim_{t \rightarrow \infty} B \simeq 1$$

when b_0 is small (as long as $\alpha + \delta < 0$), which is not unreasonable, although intuitively, one might prefer $B > 1$ under these conditions. For $n = 1$, we find that $\lim_{t \rightarrow \infty} B \simeq 1$ when $\alpha + \delta < -1$ and $b_0 \ll 1$, which is not reasonable; when the background is known to be small, even one event in the signal bin strongly favors the signal hypothesis. To prevent this pathology in the $n = 1$ case, we must require $-1 < \alpha + \delta < 0$; when we have (for $b_0 \ll 1$)

$$\lim_{t \rightarrow \infty} B = \frac{\Gamma(-\alpha - \delta)\Gamma(\alpha)b_0^{1+\alpha+\delta}}{\Gamma(-\delta)\Gamma(1 + \alpha + \delta)}$$

where, by adjusting α and δ , we can still get a wide range of behavior.

3.7 behavior with small background uncertainty

For constant y/t (estimated background) we want B to decrease monotonically as y increases. This does not always hold; the following table shows a case with 10 observed events and an estimated background of $y/t = 1$ in which B reaches a minimum at a background uncertainty of $\sim 5\%$, and increases as the background uncertainty drops further.

n	y	t	α	β	δ	B
10	1	1	0.9	0.9	-1	3.84e-2
10	10	10	0.9	0.9	-1	7.20e-5
10	100	100	0.9	0.9	-1	7.58e-6
10	1000	1000	0.9	0.9	-1	6.88e-6
10	10000	10000	0.9	0.9	-1	7.89e-6
10	100000	100000	0.9	0.9	-1	8.86e-6
10	1000000	1000000	0.9	0.9	-1	9.66e-6
10	∞	∞	0.9	0.9	-1	1.27e-5

This behavior is clearly pathological. Extensive numerical checking indicates that to avoid this pathology, one must have $-\alpha - \delta \geq 1/2$ (numerically the β parameter does not influence this behavior). Acceptable behavior is illustrated here:

n	y	t	α	β	δ	B
10	1	1	0.5	0.5	-1	4.68e-2
10	10	10	0.5	0.5	-1	1.24e-4
10	100	100	0.5	0.5	-1	1.28e-5
10	1000	1000	0.5	0.5	-1	9.42e-6
10	10000	10000	0.5	0.5	-1	9.12e-6
10	100000	100000	0.5	0.5	-1	9.095e-6
10	1000000	1000000	0.5	0.5	-1	9.0920e-6
10	∞	∞	0.5	0.5	-1	9.0916e-6

3.8 behavior when $B > 1$

We next investigate the behavior in the regime where the Bayes factor favors the $s = 0$ hypothesis. First we look at the case where $y = 1$, $t \ll 1$, and n is small (i.e. $n \ll 1/t$). This corresponds to a large background estimate with a large uncertainty, where we don't expect either hypothesis to be strongly favored. The Bayes factor is $(\alpha + \beta)/\beta$ to high precision in this case, which favors the $s = 0$ hypothesis. By adjusting α and β we can make B quite large in this case, but that would seem undesirable; with essentially no information about rate of background, we don't want the null hypothesis to be strongly favored. We therefore, somewhat arbitrarily, require that $\beta \geq \alpha$ so that $B \leq 2$ in this limit.

We can also look at a case where the observed n and y are consistent with $s = 0$ to high precision. At smaller values of $-\delta$ the behavior shows slow growth of B with increasing $n = y$, as seen here:

n	y	t	α	β	δ	B
1	1	1	0.5	0.5	-1	1.41
10	10	1	0.5	0.5	-1	2.29
100	100	1	0.5	0.5	-1	3.94
1000	1000	1	0.5	0.5	-1	6.93
10000	10000	1	0.5	0.5	-1	12.27
100000	100000	1	0.5	0.5	-1	21.80
1000000	1000000	1	0.5	0.5	-1	38.76

Larger values of $-\delta$ lead to much more rapid growth of B :

n	y	t	α	β	δ	B
1	1	1	1.5	1.5	-2	1.30
10	10	1	1.5	1.5	-2	2.99
100	100	1	1.5	1.5	-2	11.77
1000	1000	1	1.5	1.5	-2	58.61
10000	10000	1	1.5	1.5	-2	316.8
100000	100000	1	1.5	1.5	-2	1759
1000000	1000000	1	1.5	1.5	-2	9854

Some growth of B with increasing $n = y$, eventually reaching $B \gg 1$, is unavoidable in the Bayesian scheme—as the measurement achieves ever higher precision, one can effectively separate an ever larger fraction of the prior ensembles for the two hypotheses. The rate of growth of B , being determined by the prior, is under our control, and we prefer the lower growth rate associated with $\delta \simeq -1$. This is a prejudice, not a necessity, but it is probably shared by most physicists.

4 Summary

4.1 likelihood ratio

The likelihood ratio λ can be considered analogous to the Bayes factor B , although by definition $\lambda \leq 1$, while no such restriction applies to B . For the triple Poisson problem considered in this note,

$$-2 \ln \lambda = 2n \ln n + 2y \ln(y/t) - 2(n+y) \ln((n+y)/(t+1))$$

is approximately distributed as a χ^2 , in this case with 1 degree of freedom, approximately independent of the true value of the background b , provided that $b \gg 1$. Significance $\geq 5\sigma$ in this approximation then requires $\sqrt{-2 \ln \lambda} \geq 5$, equivalent to $\lambda < \sim 3.7 \times 10^{-6}$. See [8] for more discussion of this approach.

Note that λ is not a p -value; the corresponding p -value for a 5σ excursion (of either sign) is $\sim 5.7 \times 10^{-7}$.

4.2 Bayes factor

The free parameters in our prior $\pi(\rho)$, α , β , and δ , need certain constraints to avoid pathologies or unwanted behavior. The ones described above reduce to

- $0 < \alpha \leq \beta$
- $1/2 \leq -\alpha - \delta < 1$
- $1/2 < -\delta \leq \sim 1.5$

There is no guarantee of reasonable behavior everywhere in this range; in particular, values of α too close to zero should be avoided. Choosing $\alpha \simeq \beta$ seems reasonable.

The calculationally simplest choice satisfying these constraints is $\alpha = \beta = 1/2$, $\delta = -1$, when the Bayes factor is given by

$$B = \frac{\sqrt{\pi}\Gamma(y+1)(1+t)^{-1/2}}{\Gamma(y+1/2)F(-n+1, 1/2; y+1; -t)}$$

where F , when $n > 0$, reduces to a simple polynomial in t of degree $n - 1$. When $n = 0$

$$B = \frac{\sqrt{\pi}\Gamma(y+1)}{\Gamma(y+1/2)F(y, 1/2; y+1; t/(t+1))}$$

can be used; the power series should converge relatively quickly. When $n = 0$, $y = 1$, it reduces to $B = 1 + 1/\sqrt{t+1}$.

In the next table we show $\sqrt{-2 \ln \lambda}$ at approximately constant B for increasing n and y (t fixed). The general trend is that after reaching a minimum, $\sqrt{-2 \ln \lambda}$ increases slowly as n and y increase for constant B .

n	y	t	α	β	δ	B	$\sqrt{-2 \ln \lambda}$
6	2	20	0.5	0.5	-1	1.06e-5	5.27
12	27	20	0.5	0.5	-1	1.00e-5	5.25
13	33	20	0.5	0.5	-1	9.74e-6	5.25
20	86	20	0.5	0.5	-1	9.92e-6	5.24
50	415	20	0.5	0.5	-1	1.02e-5	5.25
100	1102	20	0.5	0.5	-1	1.03e-5	5.26
1000	16730	20	0.5	0.5	-1	1.06e-5	5.34
10000	189050	20	0.5	0.5	-1	1.08e-5	5.44

The consequence is that, if one uses a particular fixed cutoff value of B for “discovery”, there is no corresponding significance p -value, but there is a corresponding maximum p -value. In the case shown, $B \simeq 10^{-5}$ implies a p -value based significance $\geq \sim 5.24\sigma$.

5 Conclusions

1. The choice of (objective) priors in the calculation of the Bayes factor is a delicate issue; priors successful for parameter estimation may nevertheless prove pathological when applied to calculating the Bayes factor. In my judgment, in high energy physics there is insufficient experience in selecting priors for Bayes factor calculations, the consequences of which require extensive scrutiny, and is better left in the hands of professional statisticians.
2. It is essential that any prior that does not appear in both the numerator and denominator of the Bayes factor calculation be proper; the use of an improper prior in the Bayes factor calculation can only succeed if it is present under both hypotheses. This is in contrast to Bayesian parameter estimation, where improper priors often lead to proper posteriors.
3. The Bayes factor's exact independence of the choice of prior for the nuisance parameters μ and ϵ common to both hypotheses in this example is fortuitous, but one can hope for at least approximate independence in most cases.
4. In this example, the calculation of the Bayes factor is arguably actually easier than the calculation of the p -value, once the difficult issue of choosing a prior is settled. This is clearly not always true, but in more intractable cases, the likelihood ratio might be used as the main term in an approximation of the Bayes factor in an approach known as the Bayesian Information Criterion (BIC).
5. The Bayes factor can be viewed as a *qualitative* check of a p -value calculation, or vice versa; a *qualitative* disagreement between conclusions derived from a Bayes factor and a p -value in a particular problem necessitates deeper scrutiny of both calculations. Quoting from Jeffreys' discussion of the relative performance of Bayes factors (Jeffreys' "significance tests", the approach he greatly preferred) vs p -values (as employed by Fisher) in hypothesis testing:

I have in fact been struck repeatedly in my own work, after being led on general principles to a solution of a problem, to find that Fisher had already grasped the essentials by some brilliant piece of common sense, and that his results would be either identical with mine, or would differ in cases where we should both be very doubtful. As a matter of fact I have

applied my significance tests to numerous applications that have also been worked out by Fisher's, and have not yet found a disagreement in the actual decisions reached. ([3] p 393)

Although it is hoped that Bayes factors may eventually become useful in high energy physics, providing second opinions on questions of significance, before issues of prior selection become routine it is difficult to recommend the use of Bayes factors in anything but the simplest of contexts.

6. In the example of this note, a common scenario in HEP, $B \simeq 10^{-5}$ corresponds very roughly to $\geq 5\sigma$ significance (where B is defined as the factor in favor of the $s = 0$ hypothesis). But a more precise statement is not possible because of the remaining freedom in choice of prior for s .

References

- [1] Joel Heinrich, *Review of the Banff Challenge on Upper Limits*, in proceedings of *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*, Geneva, Switzerland, June 2007, E. H. B. Prosper, L. Lyons and A. De Roeck, eds., CERN Yellow report, cernrep/2008-001, p 125. doc.cern.ch/yellowrep/2008/2008-001/p125.pdf
- [2] P. C. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, (Cambridge University Press, Cambridge, 2005).
- [3] Harold Jeffreys, *Theory of Probability*, 3rd ed., (Oxford University Press, Oxford, 1961), §5.15, p 267.
- [4] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, Applied Mathematics Series, vol 55, Washington, National Bureau of Standards, 1964; reprinted by Dover Publications, New York, 1968.
- [5] Luc Demortier, *P-Values and Nuisance Parameters*, in proceedings of *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*, Geneva, Switzerland, June 2007, E. H. B. Prosper, L. Lyons and A. De Roeck, eds., CERN Yellow report, cernrep/2008-001, p 23. doc.cern.ch/yellowrep/2008/2008-001/p23.pdf
- [6] J. Heinrich and L. Lyons, "Systematic Errors", in *Annual Review of Nuclear and Particle Science*, Vol. 57, Palo Alto, Annual Reviews, 145 (2007) arjournals.annualreviews.org/doi/pdf/10.1146/annurev.nuc1.57.090506.123052

- [7] James Berger, *A Comparison of Testing Methodologies*, in proceedings of *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*, Geneva, Switzerland, June 2007, E. H. B. Prosper, L. Lyons and A. De Roeck, eds., CERN Yellow report, cernrep/2008-001, p 8.
doc.cern.ch/yellowrep/2008/2008-001/p8.pdf

- [8] Wolfgang A. Rolke and Angel M. Lopez, *Testing for a Signal*, in proceedings of *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*, Geneva, Switzerland, June 2007, E. H. B. Prosper, L. Lyons and A. De Roeck, eds., CERN Yellow report, cernrep/2008-001, p 34.
doc.cern.ch/yellowrep/2008/2008-001/p34.pdf