Have We Seen a Signal Yet? A Necessary Condition for Claiming Discovery

Michael Woodroofe and Bodhisattva Sen
The University of Michigan and Columbia University

March 19, 2013

Abstract

We consider experiments in which events may be either background or signals, and suppose that auxiliary variables have been attached to the events, providing partial information about their nature. The question of interest is whether any signal events are present. We adopt a Bayesian approach and derive an upper bound for the probability that a signal event has been observed that is independent of the prior distribution within broad limts.

Key words and phrases: Bayesian tests; inequalities; marked Poisson variables.

1 Marked Poisson Variables

Consider events, like neutrino oscillations, to which marks (auxiliary variables) are attached–for example, energy levels. The events may either be signal events or background, and the marks provide partial information about their nature. The problem is to determine whether any signal events have been observed. The problem is statistical but of current interest in high energy physics. [**Need more detail here**].

To formalize the statistical problem, we introduce random variables N, $J_1, J_2, \cdots$ and $X_1, X_2, \cdots$. Here N represents the total number of events observed and is supposed to have a Poisson distribution with mean $b + s$, where b and s are the expected values of the (unobserved) total number of background and signal events. The variable J_i is 0 if the i^{th} event is a background event and 1 if it is a signal; J_i takes on the values 0 and 1 with probabilities $b/(b+s)$ and $s/(b+s)$ and is not observed directly. However, X_i is observed and is supposed to have one probability density or mass function g, say, for a background event ($J_i = 0$) and another probability density or mass function

h, say, for a signal event ($J_i = 1$) and, thereby, provides partial information about J_i. Observe that $M = J_1 + \cdots + J_N$ and $N - M$ are the (unobserved) total number of signal and background events. The probability model is completed supposing that $N, (J_1, X_1), (J_2, X_2), \cdots$ are mutually independent. It can then be shown that M and $N - M$ are independent Poisson variables with means s and $b - s$ and that $X_1, X_2, \cdots$ have common (overall) density

$$\frac{bg(x) + sh(x)}{b + s}, \tag{1}$$

We suppose throughout that b and g are known and that h is known to belong to a parametric family, say $h = h_t$, where t takes values in a subset of $\mathcal{T}$ of a Euclidean space $\mathbb{R}^d$. Thus, s and t are unknown parameters. In support of these assumptions, we observe that b and g may be estimated from off-line experiments and that physical models may suggest the form of h. [**Need more here**] We call t a *nuisance parameter* below, because it complicates the discovery question.

The derivations are less cumbersome for discrete X_i, and we will consider only this case below, although our results remain valid for X_i that have a density or a mixed distribution. Further, we use $P_{s,t}$ to denote probability computed under the assumption that s is the value of the expected signal and t is the value of the nuisance parameter; and we write $\mathbf{X}$ for $(X_1, \cdots, X_N)$. Thus,

$$P_{s,t}[N = n,\ \mathbf{X} = \mathbf{x}] = \frac{e^{-(b+s)}}{n!} \prod_{i=1}^{n} [bg(x_i) + sh_t(x_i)]. \tag{2}$$

From a statistical perspective, (2) is just the likelihood function, $L(s, t|n, \mathbf{x}) = P_{s,t}[N = n, \mathbf{X} = \mathbf{x}]$. Observe that $L(s, t|n, \mathbf{x})$ does not depend on t when $s = 0$, and let $L_0(n, \mathbf{x})$ denote the common value, $L(0, t|n, \mathbf{x}) = L_0(n, \mathbf{x})$. Then

$$L(s, t|n, \mathbf{x}) = L_0(n, \mathbf{x}) \prod_{i=1}^{n} \left[1 + \frac{s}{b} r_t(x_i)\right] e^{-s}, \tag{3}$$

where $r_t(x) = h_t(x)/g(x)$.

The problem considered here is determining whether there is a positive signal, sometimes called looking for a needle in a Haystack in cases, like the Higgs, where the signal is small compared to the background. In the case of the Higgs, there is interest in significance at the 5σ level, roughly $\alpha = 10^{-6}$. The conventional formulation is as a test of $H_0 : s = 0$, though we prefer an alternative formulation, described below. We adopt a Bayesian approach and derive a necessary condition for claiming discovery, (10) below, that is independent of the prior, within broad limits. This bound is offered as a reality check on claims of a discovery. An example may illustrate its usefulness.

Example. The data in Figure 1 below were simulated from a model in which the distribution of the marks in normal with mean -1 for background events, $+1$ for signal

events, and unit standard deviation for both. For this model the locally most powerful test of $s = 0$ rejects for large values of

$$Z = \frac{\sum_{i=1}^{N}[r(X_i) - 1]}{\sqrt{b(e^4 - 1)}},$$

which has mean 0, unit standard deviation, and an asymptotic normal distribution as $b \to \infty$ under the null hypothesis.

For the data displayed in Figure 1, $Z = 5.3975...$ and the approximate P-value is 3.3789×10^{-8}, using normal approximation. That is, the data are significant at the 5σ level. However, the lower bound on the probability of on-discovery is $.00515\cdots$, a small value but far short of the 5σ criterion. There are (at least) two factors that affect the this discrepancy. One is the difference between significance levels and posterior probabilities, described by Berger and Sellke [1]. Another is the use of normal approximation far out in the tails of the distribution.

2 The Testing Problem: A Bayesian View

In a Bayesian formulation the unknown parameters are regarded as random variables or vectors, and it is convenient to have a notation for these random parameters. Thus, let S be a random variable and T be random vector with a joint apriori distribution. We will suppose that S and T are independent, apriori, with distribution functions Ξ and ρ. In the Bayesian formulation $P_{s,t}$ becomes the conditional probability given $S = s$ and $T = t$. Thus, denoting probability in the Bayesian model by P (with no sub or superscripts),

$$P[S \le u,\ N = n,\ \mathbf{X} = \mathbf{x}] = \int_0^u \int_{\mathcal{T}} P_{s,t}[N = n,\ \mathbf{X} = \mathbf{x}] d\rho(t) d\Xi(s).$$

The Conventional Formulation. The conventional way to formulate the question is as a testing problem $H_0 : S = 0$. Letting π_0 be the prior probability that $S = 0$, and ξ the conditional density of S given $S > 0$, we find

$$\begin{aligned} P[N = n,\ \mathbf{X} = \mathbf{x}] &= P[S = 0,\ N = n,\ \mathbf{X} = \mathbf{x}] + P[S > 0,\ N = n,\ \mathbf{X} = \mathbf{x}] \\ &= \pi_0 P_0[N = n,\ \mathbf{X} = \mathbf{x}] \\ &\qquad + (1 - \pi_0) \int_0^\infty \int_{\mathcal{T}} P_{s,t}[N = n, \mathbf{X} = \mathbf{x}] d\rho(t) \xi(s) ds, \end{aligned}$$

where P_0 denotes the common value of $P_{0,t}$. Using (3), this is

$$P[N = n,\ \mathbf{X} = \mathbf{x}] = L_0(n, \mathbf{x}) \left[\pi_0 + (1 - \pi_0) Q_{\rho,\xi}(n, \mathbf{x})\right], \tag{4}$$

where

$$Q_{\rho,\xi}(n, \mathbf{x}) = \int_0^\infty \int_{\mathcal{T}} \prod_{i=1}^{n} \left[1 + \frac{s}{b} r_t(x_i)\right] e^{-s} d\rho(t) \xi(s) ds. \tag{5}$$

So, the posterior probabilty that $S = 0$ is

$$\pi_0^* := P[S = 0|N = n, \mathbf{X} = \mathbf{x}] = \frac{\pi_0}{\pi_0 + (1 - \pi_0)Q_{\rho,\xi}(n, \mathbf{x})},$$

and the posterior odds in favor of $S = 0$ are

$$\frac{\pi_0^*}{1 - \pi_0^*} = \left(\frac{\pi_0}{1 - \pi_0}\right) \frac{1}{Q_{\rho,\xi}(n, \mathbf{x})}.$$

The Bayesian test is to reject H_0 if π_0^* is sufficiently small. Unfortunately, π_0^* depends crucially on π_0. The (so called) Bayes Factor $1/Q_{\rho,\xi}(n, \mathbf{x})$ represents the amount of change in the odds. It does not depend on π_0, but only on ρ and ξ.

An Alternative Formulation. Another way to pose the discovery problem is to ask whether $M > 0$; that is, have we seen a signal event yet? As in (4),

$$\begin{aligned} P[M = 0, N = n, \mathbf{X} = \mathbf{x}] &= \int_0^\infty \int_{\mathcal{T}} \frac{b^n e^{-(b+s)}}{n!} \prod_{i=1}^n g(x_i) d\rho(t) d\Xi(s) \\ &= L_0(n, \mathbf{x})[\pi_0 + (1 - \pi_0)\mu_0], \end{aligned}$$

where $\mu_0 = \int_0^\infty e^{-s}\xi(s)ds$. Combining this with (4) and (5),

$$P[M = 0|N = n, \mathbf{X} = \mathbf{x}] = \frac{\pi_0 + (1 - \pi_0)\mu_0}{\pi_0 + (1 - \pi_0)Q_{\rho,\xi}(n, \mathbf{x})}. \tag{6}$$

The alternative test is to reject if (6) is small. In the next section, it is shown that (6) is minimized when $\pi_0 = 0$ and has an easily computable lower bound that does not depend on ξ or ρ within broad limits.

Of course, if $M > 0$, then necessarily $s > 0$; and it seems unlikely (to us) that anyone would want to claim a discovery without strong evidence that $M > 0$. In this sense, the two formulations are roughly equivalent. To the extent that the two formulations differ, we prefer the alternative, because it is closer to the data: The variable M is defined in terms the experiment; the parameter s is only defined in terms of the mathematical model for the experiment. This is very much in the spirit of [2].

Computational Issues and Some Algebra. For later reference observe that

$$\prod_{i=1}^n \left[1 + \frac{s}{b} r_t(x_i)\right] = \sum_{k=0}^n C_{n,k}(t)(\frac{s}{b})^k,$$

where

$$C_{n,k}(t) = \sum_{j_1 + \cdots + j_n = k} \prod_{i=1}^n r_t(x_i)^{j_i},$$

and the summation extends over all $(j_1, \cdots, j_n)$ for which $j_i = 0$ or 1 for each i and $j_1 + \cdots + j_n = k$. The $C_{n,k}$'s may be computed recursively using

$$C_{n,k}(t) = C_{n-1,k}(t) + r_t(x_n)C_{n-1,k-1}(t),$$

as in Roe and Woodroofe [3]. Then

$$Q_{\rho,\xi}(n,\mathbf{x}) = \sum_{k=0}^{n} \bar{C}_{n,k} \left(\frac{\mu_k}{b^k}\right) \tag{7}$$

where

$$\mu_k = \int_0^\infty s^k e^{-s} \xi(s) ds,$$

$$\bar{C}_{n,k} = \int C_{n,k}(t) d\rho(t),$$

3 Inequalities

For situations in which the signal is thought to be small, it seem reasonable to suppose that ξ is a non-increasing function. A lower bound for (6) over all $0 < \pi_0 < 1$, all non-increasing ξ, and all ρ is developed in steps below; but first an auxillary result is needed.

The Correlation Inequality. The following result is intuitive: the correlation between a non-decreasing function of a random variable and a non-increasing function of the same variable is non-positive: that is, if F is a distribution function, u is a non-decreasing function, and v is a non-increasing function for which

$$\int_{I\!R} (|u| + |v|) dF < \infty,$$

then

$$\int_{I\!R} uv dF \leq \left[\int_{I\!R} u dF\right] \left[\int_{I\!R} v dF\right]. \tag{8}$$

See, $\cdots$ for the details.

Dependence on π_0. For any ξ, (6) is minimized when $\pi_0 = 0$. This is also intuitive. To see it mathematically observe first that μ_0 is the first term in the sum (7), so that $\mu_0 < Q_{\rho,\xi}(n,\mathbf{x})$; and clearly $\mu_0 < 1$. So, the derivative of the logarithm of (6) with respect to π_0 is

$$\begin{aligned}\frac{1-\mu_0}{\pi_0 + (1-\pi_0)\mu_0} - \frac{1 - Q_{\rho,\xi}(n,\mathbf{x})}{\pi_0 + (1-\pi_0)Q_{\rho,\xi}(n,\mathbf{x})} \\ \geq \frac{1-\mu_0}{\pi_0 + (1-\pi_0)Q_{\rho,\xi}(n,\mathbf{x})} - \frac{1 - Q_{\rho,\xi}(n,\mathbf{x})}{\pi_0 + (1-\pi_0)Q_{\rho,\xi}(n,\mathbf{x})} \geq 0,\end{aligned}$$

and the minimum is attained when $\pi_0 = 0$,

Dependence on ξ. By (8) applied with $u(s) = s^k$, $v(s) = \xi(s)$, $F(s) = 1 - e^{-s}$ for $s > 0$ and, therefore, $dF(s) = e^{-s} ds$,

$$\mu_k = \int_0^\infty s^k e^{-s} \xi(s) ds \leq \left[\int_0^\infty s^k e^{-s} ds\right] \left[\int_0^\infty e^{-s} \xi(s) ds\right] = k! \mu_0.$$

So, when $\pi_0 = 0$

$$\frac{1}{P[M=0|N=n,\mathbf{X}=\mathbf{x}]} = \frac{Q_{\rho,\xi}(n,\mathbf{x})}{\mu_0} = \sum_{k=0}^{n} \frac{1}{b^k} \bar{C}_{n,k} \left(\frac{\mu_k}{\mu_0}\right) \leq \sum_{k=0}^{n} \frac{k!\bar{C}_{n,k}}{b^k},$$

which is the same value obtained from (6) by letting $\pi_0 = 0$ and formally setting $\xi(s) \equiv 1$. So,

$$P[M=0|N=n,\mathbf{X}=\mathbf{x}] \geq \frac{1}{\sum_{k=0}^{n} k!\bar{C}_{n,k} b^{-k}} \tag{9}$$

for all $0 < \pi_0 < 1$, all non-increasing ξ, and all ρ.

Dependence on ρ. The right side of (9) still depends on ρ, but may be easily bounded. Let $\bar{L}$ denote the integrated likelihood

$$\bar{L}(t|n,\mathbf{x}) = \int_0^\infty L(s,t|n,\mathbf{x})ds.$$

Then

$$\bar{L}(t|n,\mathbf{x}) = L_0(n,\mathbf{x})Q(n,\mathbf{x},t),$$

where

$$Q(n,\mathbf{x},t) = \int_0^\infty \prod_{i=1}^{n} \left[1 + \frac{s}{b} r_t(x_i)\right] e^{-s} ds = \sum_{k=0}^{n} \frac{k!C_{n,k}(t)}{b^k}.$$

Next, let $\hat{t}$ maximize the integrated likelihood function, so that

$$Q(n,\mathbf{x},\hat{t}) = \max_t Q(n,\mathbf{x},t) = \hat{Q}(n,\mathbf{x}), \text{ say.}$$

Then

$$\sum_{k=0}^{n} k!\bar{C}_{n,k} b^{-k} \leq \hat{Q}(n,\mathbf{x}) = \sum_{k=0}^{n} \frac{k!C_{n,k}(\hat{t})}{b^k},$$

in (9), and

$$P[M=0|N=n,\mathbf{X}=\mathbf{x}] \geq \frac{1}{\hat{Q}(n,\mathbf{x})} \tag{10}$$

for all $0 < \pi_0 < 1$, all non-increasing ξ, and all ρ. So, from a Bayesian perspective, a necessary condition for claiming discovery is that

$$\sum_{k=0}^{n} \frac{k!C_{n,k}(\hat{t})}{b^k} \geq \frac{1}{\alpha} \approx 10^6. \tag{11}$$

4 Remarks, References, and Acknowledgments

The derivation of (11) also leads to a bound on the Bayes factor: Since $\mu_0 \leq 1$, $Q_{\rho,\xi}(n,\mathbf{x})$ is at most the left side of (11). The derivation of (11) is patterned after Berger and Sellke [1], and the bound on the Bayes factor is a special case of their results.

The research of both authors was supported by the National Science Foundation.

References

[1] Berger, James and Tom Sellke (1987). Testing a point null hypothesis: The irreconcillability of p values and evidence. *J. Amer. Statist. Assn*, **82**, 112-122.

[2] Geisser, Seymour (1993). *Predictive Inference*. Chapman Hall.

[3] Roe, Byron and Michael Woodroofe (2000). Setting confidence belts. *Physics Review, D*, **63**, 013009-013018.