Setting Limits, Computing Intervals, and Detection

David A. van Dyk
Department of Statistics
University of California, Irvine, CA, United States

Abstract

This article discusses a number of statistical aspects of source detection, the computation of intervals and upper limits for a source intensity, and accessing the sensitivity of a detection procedure. Emphasis is placed on model diagnostics, validation, and improvement as means of avoiding odd behaviors in these procedures such as over abundant short or empty intervals. Improved model specification is viewed as a better response to systematic uncertainties, the look elsewhere effect, and general model inadequacy than simply insisting on a significance level of 5σ for source detection. We advocate reporting *both* the upper limit and the sensitivity to better represent the strength of evidence for detection and the reported source intensity. Finally, we explore the use of decision theoretic analysis to derive detection procedures, intervals, and limits in order to focus attention on the statistical properties of primary interest.

1 Introduction

Over the past 10-15 years there has been much discussion in the high energy physics community as to how best to derive statistical criterion for source detection and how best to compute intervals and limits for source intensities, see e.g., [1–3]; and the proceedings for the Phystat Conference Series (`URL: physstat.org`). This paper picks up a number of threads in this discussion from a statistical point of view and with an emphasis on encouraging adequate model specification and proper reporting of results. From my point of view the discussion has been too focused on technical properties and somewhat superficial concerns pertaining to statistical procedures. Thus, this paper explores a decision theoretic approach with the aim of focusing attention on the statistical properties most pertinent to ultimate scientific goals.

The paper is organized into five sections. In Section 2 we review the basic statistical framework for source detections and setting intervals and upper limits for the source intensity. Important in this is the clarification of a difference in nomenclature used in high energy physics and in astrophysics. In Section 3 we discuss a number of concerns that have arisen with this framework. The use of decision theoretic analysis to derive new procedures for detection and computing intervals and limits is explored in Section 4. The paper is summarized in Section 5.

2 Detection, Intervals, and Upper Limits

2.1 A Simple Poisson Model

To focus attention on the statistical issues we frame our discussion in terms of a simple detection problem involving a contaminated Poisson count. The methods and issues described are general, but the salient points are evident in this simple example. Thus, we consider the Poisson model for a source count,[1]

$$n|(\lambda_S, \lambda_B, \tau_S) \sim \text{Poisson}\Big(\tau_S(\lambda_S + \lambda_B)\Big), \tag{1}$$

where n is the source count, λ_S is the source intensity, λ_B is the background intensity, and τ_S is the source exposure time. We typically have a second background-only exposure that we model as

$$n_B|(\lambda_B, r, \tau_B) \sim \text{Poisson}(r\tau_B\lambda_B), \tag{2}$$

[1] The notation $X|Y \sim \text{Distribution}(Y)$ describes the *conditional* distribution of X *given* Y. For example, $X|Y \sim \text{Poisson}(g(Y))$ means that the conditional probability mass function of X given Y is $\exp\{-g(Y)\}g(Y)^X/X!$.

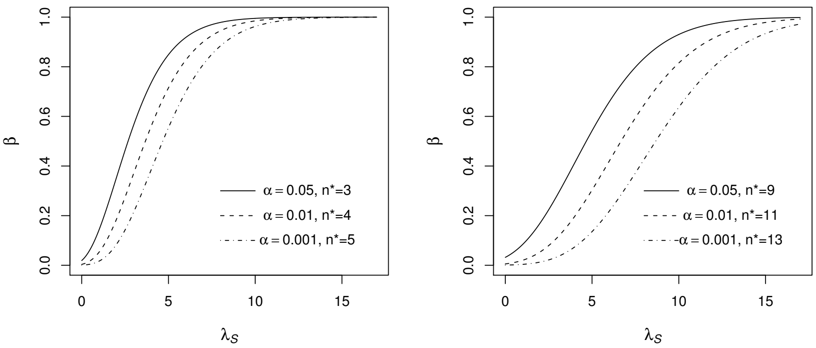


Fig. 1: The Power of the Detection Plotted as a Function of the Source Intensity, λ_S. The two panels correspond to $\lambda_B = 1$ and 5. In each panel the power is given for three values of α and their corresponding detection thresholds. The power of the detection increases with the source intensity and decreases with the background intensity. Insisting on a lower probability of a false detection (smaller α) decreases the power of the detection.

where n_B is the background count, τ_B is the background exposure time, and r is the relative area of the background and source exposures. For clarify, we sometimes assume λ_B is known. In any case, λ_S is of primary interest. We wish to determine if there is a source and if so how strong it is. Even if we cannot detect a source, we may wish to quantify how strong a possible source could be and go undetected.

A standard statistical hypothesis testing framework is used for source detection. In particular the default or *null hypothesis* states that there is no source. We assume this to be true unless we find this assumption to be at odds with the observed data, in which case we reject the null hypothesis in favor of the *alternative hypotheses* that a source is present. Formally, we write

$$H_0 : \quad \text{There is no source, i.e., } \lambda_S = 0 \tag{3}$$

$$H_A : \quad \text{There is a source, i.e., } \lambda_S > 0. \tag{4}$$

2.2 Detection

To determine whether the observed data are at odds with the null hypothesis, we first identify a *test statistic* which is a function of the data for which larger (or smaller) values correspond to stronger evidence against the null hypothesis. In our simple Poisson example, the source count, n, is an obvious choice. Having identified a test statistic, we define the *detection threshold*, $n^\star$, as the smallest value such that

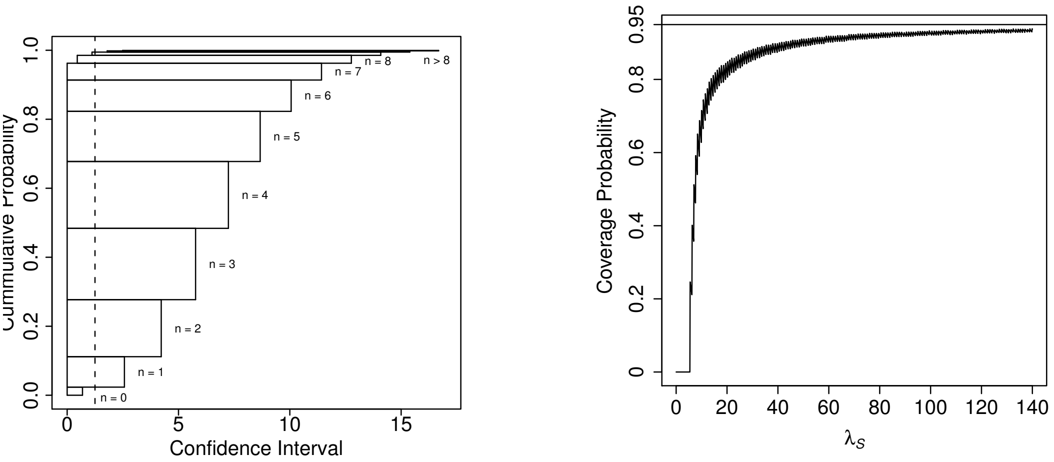
$$\Pr(n > n^\star | \lambda_S = 0, \lambda_B, \tau_S, \tau_B, r) \leq \alpha. \tag{5}$$

By conditioning on $\lambda_S = 0$ we are assuming there is no source. Under the null hypothesis the probability of a source count larger than $n^\star$ is less than or equal to the *significance level of the detection*, α. If α is set sufficiently small, and the source count is greater than $n^\star$, we conclude that there is sufficient evidence to *reject the null hypothesis* in favor of the alternative hypothesis that a source is present.

We choose a small value of α to minimizing the probability of a false detection. Of course, we can compute $\Pr(n > n^\star)$ for positive values of λ_S, in which case this becomes the probability of a true detection, which we would like to be as large as possible. The probability of a true detection depends on the the value of λ_S, is known as the *power of the detection*, and can be written

$$\beta(\lambda_S) = \Pr(n > n^\star | \lambda_S, \lambda_B, \tau_S, \tau_B, r)\,. \tag{6}$$

Note $\beta(\lambda_S = 0) \leq \alpha$ and $\beta(\lambda_S)$ is simply the probability of a detection, false or true depending on λ_S. The dependencies of the power on the source intensity and the level of the test are illustrated in Fig. 1.



(a) Sampling Distribution of a 95% Interval. The horizontal ranges of the rectangles give the confidence intervals for the given value of n, with $\lambda_B = 3.0$. Rectangle heights are the probabilities of each n and thus the probabilities of the intervals, see [2].

(b) Under Coverage. The plot shows the true coverage of 95% intervals that are only reported when a source is detected with significance level $\alpha = 0.05$ and with $\lambda_B = 5$. The true coverage is far below the nominal coverage for weak sources.

Fig. 2: Distribution and Under Coverage of Selectively Reported Confidence Intervals of [4].

2.3 Confidence Intervals, Sensitivity, Upper Limits, and Upper Bounds

A formal hypothesis test is only the first step in source detection. Whether or not there is a detection, we typically want to quantify the plausible values for the (possible) source intensity. This is certainly of interest in the event of a detection, but even in the absence of detection there is typically a non-zero probability of a *false negative*, that is, an undetected source. Formally, this probability that a source goes undetected is $1 - \beta(\lambda_S)$ and is generally expected to diminish as λ_S increases but to approach $1 - \alpha$ for λ_S near zero. (In principle $\beta(\lambda_S)$ may be discontinuous at zero or may not asymptotically approach one, but these are unusual cases.) Thus, even in the absence of a detection, a quantification of the plausible values of λ_S is of value. This quantification typically takes the form of an upper limit and/or an interval.

A frequentist *confidence interval* for λ_S aims to give the *plausible values of* λ_S. This is defined to be any interval that includes the true value of λ_S a given proportion of the time over the long run upon repetition of an experiment. Formally, we can derive an interval $\mathcal{I}(\lambda_S)$ for each value of λ_S, such that

$$\Pr(n \in \mathcal{I}(\lambda_S)|\lambda_S) \geq 95\%, \tag{7}$$

where 95% is the *confidence level* and can be replaced by any desired level. Upon observing a particular value of n_{obs} of n, a frequency confidence interval can be constructed as

$$\{\lambda_S : n \in \mathcal{I}(\lambda_S)\}. \tag{8}$$

Here we avoid the issue of *nuisance parameters*, such as λ_B. The probability in Equation 7 clearly depends on λ_B and thus so do the intervals $\mathcal{I}(\lambda_S)$ which complicates the construction of the confidence interval in Equation 8. Although this is an important issue, it is not central to our discussion, and we will simply fix λ_B at some known value when computing confidence intervals. Fig. 2(a), for example, illustrates the frequency properties of a Garwood's (1936) choice of interval for λ_S.

The upper end point of a one-sided confidence interval is called an *upper limit* by physicists (or an *upper bound* by astronomers). This is the largest plausible value of the source intensity consistent with the observation. Fig. 2(a) illustrates how one sided confidence intervals arise when n is relatively small.

In astronomy, an *upper limit* is used to quantify the source intensity of a possible, but undetected source. In particular, to an astronomer an upper limit is *the maximum intensity that a source can have*

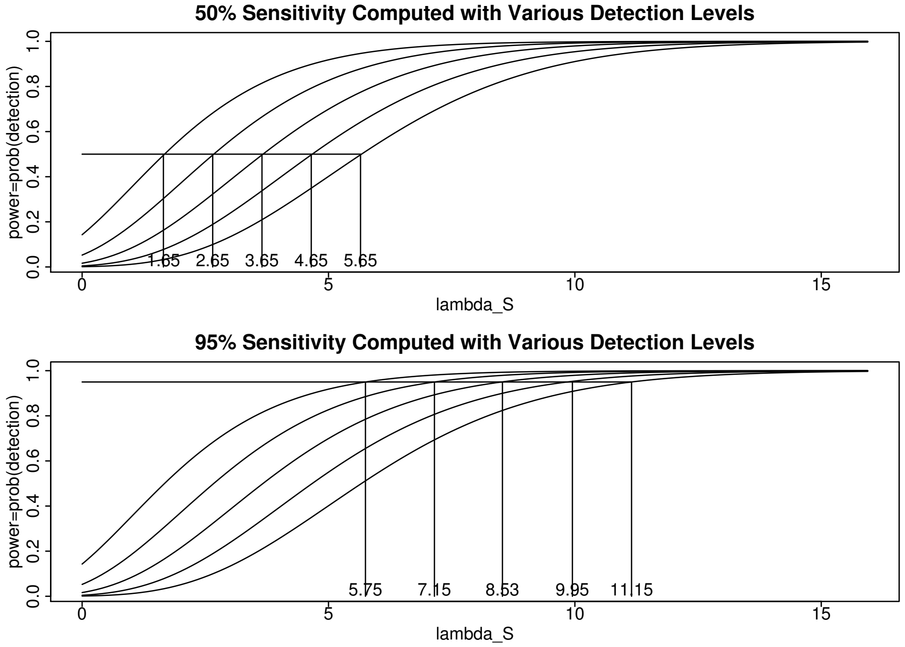


Fig. 3: Effect of α and $\beta_{\min}$ on the Upper Limit. The five curves in each panel give the probability of detection, $\beta(\lambda_S)$ for each of five values of the significance level, α, from left to right: 0.143, 0.053, 0.017, 0.005, and 0.001. When computing sensitivities we derive the minimum value of λ_S that has at least a probability of $\beta_{\min}$ of being detected. This is done for $\beta_{\min} = 0.50$ in the first panel and 0.95 in the second. The sensitivity of the detection increase as $\beta_{\min}$ increases and as α decreases.

without having at least a probability of $\beta_{\min}$ of being detected under an α-level detection threshold, or conversely, *the smallest intensity that a source can have with at least a probability of $\beta_{\min}$ of being detected under an α-level detection threshold*, see [5]. Physicists generally refer to this as the *sensitivity* of the detection. We will use the term "sensativity" from now on. Computing the sensitivity requires two probability calculations. The detection threshold is computed with the probability calculation in Inequality 5 and the probability of detection is computed using Equation 6. This is illustrated in Fig. 3.

The sensitivity of the detection is analogous to a sample size in that they both quantify the strength of an experiment. Larger sample sizes correspond to more powerful experiments that can detect weaker signals. Likewise smaller (i.e., better) sensitivities indicate a more powerful observation: any source with intensity greater than the sensitivity is expected to be detected (as calibrated by α and $\beta_{\min}$). The sensitivity directly quantifies the power in terms of the quantity of primary interest: the source intensity.

In a typical statistical power calculation, we find the minimum exposer time, τ_S, by solving Equation 6 so that the probability of detection achieves a minimum value for a given λ_S. For example, we might want to find the minimum exposure time so that $\beta(\lambda_S = 2) \geq 0.90$ if we want to be sure there is at least a 90% chance of detecting a source with intensity equal to two counts per unit time. The sensitivity of the detection is found by solving the same equation, but for λ_S with τ_S fixed. It is important to notice that all of these calculations can be done *before the observation is made*. Like power, the sensitivity does not depend on the data and can be computed in advance.

3 Addressing Concerns (Please Forgive my Soap Box!)

3.1 What Should be Reported?

A typical procedure for source detection in astronomy involves reporting different quantities depending on whether the source is detected [5]. When there is a detection astronomers often (i) report a detection and (ii) report a confidence interval for λ_S. When there is not a detection astronomers often (i) report no detection and (ii) report a detection sensitivity for λ_S. Similarly, with power-constrained limits, the data-dependent upper limit is only reported if it is greater than the sensitivity of the detection, otherwise the data-independent sensitivity is reported, see, e.g., [6]. Deciding whether or not to report an interval (or limit) *based on the data* alters its frequency properties [5,7]. This is illustrated in Fig. 2(b) which reports the frequency coverage of intervals that are only reported in the case of a detection. For small values of λ_S, the coverage can be far below its nominal value. Unfortunately, frequency properties depend on what you would have done, had you had a different data set.

To eliminate the coverage problems described in Fig. 2(b) and to provide a more complete summary of what was learned from the observation, [5] proposes that we *always report*

1. whether the source was detected,
2. a confidence interval for the source intensity (which may be a one-sided upper limit), and
3. the sensitivity of the detection, in order to quantify the strength of the experiment.

This is in contrast to both the power-constrained limit that report the larger of the sensitivity and the upper limit and to CL_S [8] that alters the upper limit in order to produce a smoothed version of the power-constrained limit [9]. Both of these procedures sacrifice frequency properties and lack a clear probabilistic interpretation. By reporting both the upper limit and the sensitivity, we provide both the largest value of λ_S consistent with the data (the upper limit) and the smallest value that we have sensitivity to detect. Reporting both the upper limit and the sensitivity is certainly more informative than reporting either max(upper limit, sensitivity) or a smoothed version of this maximum.

3.2 Short or Empty Confidence Intervals

One particular concern regarding available methods is the possibility that frequency-based intervals may be empty or very short. The former case is generally disconcerting and the later is interpreted by some users as implying an exaggerated experimental sensitivity. In my view this stems for a simple misunderstanding of the proper interpretation of the frequency-based intervals. Recall that a (say) 95% frequency-based interval is simply an interval constructed so that there is a 95% probability that an experiment conducted as formalized by the probabilistic model will result in an interval that contains the true value of λ_S. Fig. 2(a) illustrates that the same experiment sometimes produces relatively short and sometimes produces relatively long intervals. The sensitivity of an experiment, however, does not depend on the observed count. In the example in Fig. 2(a), the sensitivity is the same regardless of whether we observe $n = 0$ and obtain a short interval, or observe $n = 8$ and obtain a long interval.

Another difficulty is a tendency to interpret the *pre-data probabilities* associated with frequency intervals as *post-data probabilities*. A 95% interval will produce intervals that contain the true value of λ_S 95% of the time when observations are generated under the model, regardless of the true value of λ_S. Such a procure can produce empty intervals, so long as they are produced less that 5% of the time and overall at least 95% of the intervals contain the true value. (Of course the empty intervals may be wasteful!) This is not to say that an empty—or any other particular—interval has a 95% chance of containing the true value. An empty interval certainly does not contain the true value of λ_S, regardless of the frequency probability of the interval. Although our intuition leads us to interpret these probabilities in a post-data manner, frequency-based probabilities say nothing about the properties of a particular interval. Bayesian methods are better suited to quantifing post-data probabilities. The precise nature of frequency probabilities may be appealing, but precise probabilities are not necessary relevant probabilities.

Under the construction described in Section 2.3, we can further interpret the intervals as reporting values of λ_S that are *consistent with the observation*, where "consistent" is calibrated by the probability level associated with the interval. Short or empty intervals simply mean that there are few or no values of λ_S that are consistent with the observations. As illustrated in Fig. 2(a), very short intervals are possible, but are expected to be rare. Depending on how the interval is constructed, the same can be said for empty intervals. If empty or short intervals (relative to the sensitivity) are common, it is a clear indication that the probabilistic model used to describe the observation is inadequate—regardless of the strength of the *subjective prior belief in the underlying model.* Model checking, validation, and improvement are standard components of any statistical analysis. I expect far more would be gained by focusing on model improvement rather than on statistical properties of a particular statistical procedure.

3.3 5σ

It has become standard to require $\alpha = 1/1.7 \times 10^6$ for a detection in high energy physics, corresponding to the probability that a standard normal variable exceeds five standard deviations from its mean. This corresponds to a false positive rate of one in 1.7million experiments. Of course, the motivation is not to keep the false detection rate this low, but to attempt to account for other concerns such as the look elsewhere effect [3, 10], calibration and/or systematic errors, and statistical error rates that are not well calibrated due to general model misspecification [3, 11]. Unfortunately, reducing α does not really address these concerns. We do not know the actual effects of systematics and the look elsewhere effect on the final analysis. They likely induce both increased bias and variance. Reducing α does not address bias at all and is a completely uncalibrated response to variance. Even in the absence of these problems, statistical procedures are not well calibrated at such extreme depths in the tails of the sampling distributions, which are typically based on asymptotic approximations. Computing extreme tail probabilities poses its own challenges in all but the simplest cases [12]. Taken together these concerns lead us to conclude that we have no idea what the probability of a false detection is—the procedure itself is wholly uncalibrated.

The difficulty here is similar to what leads to over-abundant empty or narrow confidence intervals: model misspecification. The solution is not to crank down the value of α, but rather to directly deal with systematics, calibration, the look elsewhere effect, and general model misspecification. Model checking and improvement are the key to better statistical properties of detection procedures, intervals, and limits [2, 13]. Hiding unrealistic assumptions and using *ad hoc* fixes (such as using a 5σ detection criterion) do not address the root problems, but do make evaluating their effects more difficult. Calibration, systematics, and the look elsewhere effect must be modeled directly. Reasonable model specification is far more important than the detailed properties of a statistical procedure or the choice of a Bayesian, Frequentist, or other procedure. The ultimate goal is honest frequency error rates and/or a calibrated Bayesian procedure, both of which depend absolutely on careful model specification.

4 A More Coherent Approach?

4.1 Hypothesis Testing in High Energy Physics

Source detection in high energy physics is often conducted using a more involved hypothesis-testing procedure than is described in Sections 2–3. In addition to testing the hypotheses in Equations 3–4, a second hypothesis test is often conducted in tandem that interchanges the roles of the null and alternative hypotheses, see [14]. Rather than under the default assumption of no source, a second "detection threshold" is computed under the assumption that there is a source and the significance test is conducted treating the original alternative hypothesis as the null hypothesis and treating the original null hypothesis as the alternative hypothesis. (For clarity, we continue to use H_0 for the hypothesis of no source and H_A for the hypothesis that there is a source. In the reversed formulation of the significance test, we assume H_A when computing the second detection threshold, $n^\star_A$, in analogy to Equation 5.)

This reversed formulation of the hypothesis test is motivated by a well-known challenge associated

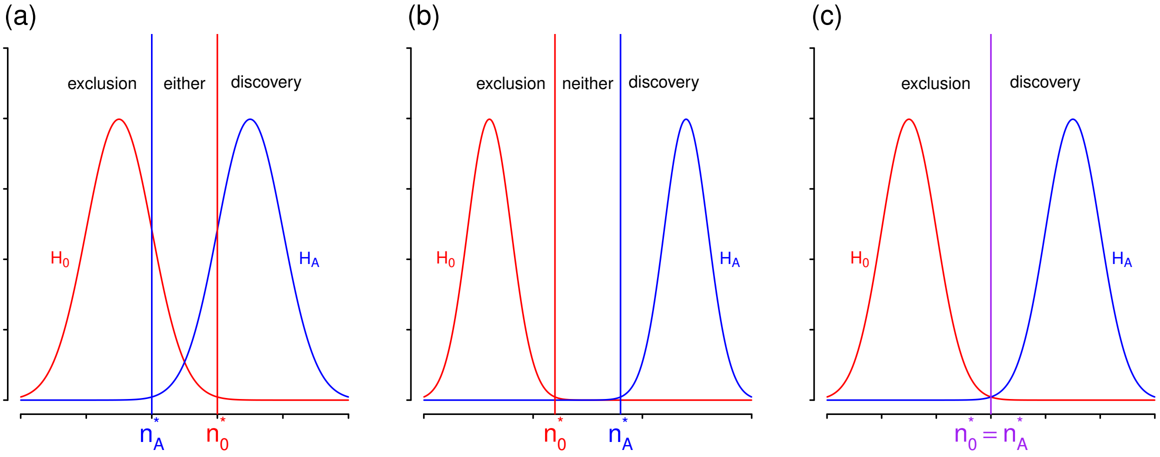


Fig. 4: Combining the Original and Reversed Significance Tests. The two curves in each panel depict the distribution of the test statistic under H_0 (left, red) and H_A (right, blue). Although we use the notation of our running example, here we assume that both distributions are fully specified, i.e., that they do not depend on any unknown parameters. The "detection" thresholds are denoted by $n_0^\star$ and $n_A^\star$, where $n_0^\star$ is the 1-α percentile of the distribution of the test statistic under H_0 and $n_A^\star$ is the α percentile under H_A. The three panels give the decision regions under three scenarios: (a) $n_0^\star > n_A^\star$, (b) $n_0^\star < n_A^\star$, and (c) $n_0^\star = n_A^\star$. We accept H_0, but reject H_A if $n < \min(n_0^\star, n_A^\star)$; reject H_0, but accept H_A if $n > \max(n_0^\star, n_A^\star)$; reject both H_0 and H_A if $n_0^\star < n < n_A^\star$; and accept both H_0 and H_A if $n_A^\star < n < n_0^\star$. Notice that in each of the scenarios, at most three of the four decisions is possible.

with *model selection*: a model being the better of two at explaining the data does not mean that it is an adequate model. In the context of hypothesis testing, rejecting the null hypothesis indicates that the hypothesis is inadequate for explaining the data, at least in the dimension quantified by the test statistic. This alone, however, is not enough for us to conclude that the alternative hypothesis *is* adequate. There are other possibilities besides the model given in Equations 1–2 with $\lambda_S = 0$ and with $\lambda_S > 0$. The reversed hypothesis test aims to identify evidence that $\lambda_S > 0$ is inadequate as well. Of course, all hypothesis tests look for evidence in the dimension specified by the test statistic, so the interplay of the original and the reversed hypothesis tests depends intimately on the two choices of test statistics.

The decision in the original hypothesis testing framework involves either *accepting* H_0 or *rejecting* H_0. When we conduct both the original and the reversed hypothesis test, each test has these two possible outcomes, leading to a total of four possibilities:

exclusion: accept H_0 and reject H_1,

discovery: reject H_0 and accept H_1,

no decision: accept both hypotheses (*either* is possible), or

excluding both: reject both hypotheses (*neither* is possible).

As illutrated in Figure 4, in any particular situation only one of "no decision" and "exclude both" is possible, depending on the ordering of the detection thresholds for the two hypothesis tests.

While it is completely standard to use model diagnostics and checking to evaluate the adequacy of any statistical model, formal symmetric testing of H_0 and H_A in this way is unusual, if not unique to high energy physics. Inverting a significance test to form confidence intervals or upper limits is a related and very common technique. This involves treating each possible value of the parameter as a null hypothesis and compiling the interval as the set of parameter values that are not rejected at a given α-level. An additional complication arrises in high energy physics in that different significance levels are used for the original and the reversed significance tests, typically 5σ and 2σ, respectively. In the following section we employ a decision theoretic approach to analyze the use of such symmetric testing.

Table 1: Loss Functions. Table (a) gives a detailed loss function for the six possible errors if we assume either H_0 or H_A is true. To simplify calculations, Table (b) gives a loss function where the cost of all errors except a false detection are equal.

(a)

	Decision			
Truth	exclusion	discovery	no decision	exclude both
H_0	0	C_{01}	C_{0e}	C_{0n}
H_A	C_{10}	0	C_{1e}	C_{1n}

(b)

	Decision			
Truth	exclusion	discovery	no decision	exclude both
H_0	0	C	c	c
H_A	c	0	c	c

4.2 A Decision Theoretic Approach: Loss, Risk, and Bayes Risk

Although concerns about detection procedures are often expressed in terms of detailed observations about the character of procedures under certain circumstances (e.g., the upper limit may increase as n decreases), a desire for strict adherence to frequency properties (e.g., the "Goldilocks effect": coverage should be above a minimum, but no more than the minimum); and apprehension about Bayesian methods and their prior distributions, e.g., [1], ultimately we are primarialy concerned with rates of detection errors and ensuring that intervals and limits do a good job of capturing the true source intensities. In this section, we discuss a *decision theoretic analysis* that allows us to directly optimize a detection procedure in terms of the quantities of ultimate interest.

We begin with a loss function that quantifies the cost of the possible errors in a significance test with the four possible decision: "exclusion", "discovery", "no decision", and "exclude both". With four possible decisions there are more possible errors than the "false detection" and "false negative" of a standard significance test, see Table 1(a). While it can be argued that "no decision" is not an "error" regardless of the truth, this decision is clearly less desirable than a true exclusion or a true discovery. In this regard it is appropriate to assign a non-zero loss to this decision, even if it is not an "error". A more complete table would include a third row, "Truth = Neither" to capture the possibility that neither H_0 nor H_A holds. We avoid this possibility because the necessary probability calculations are arbitrary when no true model is specified. In Table 1(a), C_{01} is the cost of the most troubling error, a false positive. The costs of the all other errors are likely significantly smaller than C_{01}. The loss function in Table 1(b) quantifies this by setting the cost of all other errors to $c \ll C = C_{01}$. That is, for simplicity we assume that all errors except a false detection have an equal cost that is dominated by the cost of a false detection. Finally we assume that $C + c = 1$; this is simply a choice of scale for the loss function.

Given detection thresholds, $n_0^\star$ and $n_A^\star$, we compute the *risk*, which is the expected loss, under H_0,

$$\text{Risk}(n_0^*, n_A^*|H_0) = C\,\Pr[n > \max(n_0^*, n_A^*)|H_0] + c\Big\{\Pr[n_0^* < n < n_A^*|H_0] + \Pr[n_A^* < n < n_0^*|H_0]\Big\}$$

and under H_A,

$$\text{Risk}(n_0^*, n_A^*|H_1) = c\,\Pr[n > \min(n_0^*, n_A^*)|H_1] + c\Big\{\Pr[n_0^* < n < n_A^*|H_1] + \Pr[n_A^* < n < n_0^*|H_1]\Big\}.$$

Our goal is to find $n_0^\star$ and $n_A^\star$ to minimize the risk. The *Bayes risk* averages $\text{Risk}(n_0^*, n_A^*|H_0)$ and $\text{Risk}(n_0^*, n_A^*|H_A)$ using a probability of H_A, denoted by π,

$$\text{Bayes Risk}(n_0^*, n_A^*|\pi) = (1-\pi)\,\text{Risk}(n_0^*, n_A^*|H_0) + \pi\,\text{Risk}(n_0^*, n_A^*|H_A).$$

To minimize the Bayes risk, we make a simplifying assumption that the test statistic has a continuous distribution with probability density function f_0 under H_0 and f_A under H_A. This is not the case in the Poisson model, where n is a count. Under this assumption the Bayes risk is minimized either when

$$C = \frac{(1-\pi)f_0(n_0^*) + \pi f_A(n_0^*)}{2(1-\pi)f_0(n_0^*) + \pi f_A(n_0^*)} = \frac{(1-\pi)f_0(n_A^*) + \pi f_A(n_A^*)}{2(1-\pi)f_0(n_A^*) + \pi f_A(n_A^*)}$$

or at a point where the Bayes risk is not differentiable, $n_0^* = n_A^*$. Thus the optimal choice of $n_0^\star$ and $n_A^\star$ occurs when $n_0^\star = n_A^\star$, with the particular optimal value of $n_0^\star = n_A^\star$ determined by C and c. This corresponds to the standard detection setup in that there are only two possible decisions, see Fig. 4(c). This result depends on the simple loss function given in Table 1(b) and would be different if different costs were assigned to a false exclusion and the "no decision" and "exclude both" decisions under H_0 and/or H_A. Of course quantifying the relative costs of the various errors in Table 1 is not an easy task.

The result can be understood by referring to Fig. 4(a). Suppose we fix $n_0^\star$ and adjust $n_A^\star$ with the aim of decreasing the risk under H_0. Increasing $n_A^\star$ increases the probability of the correct (zero cost) decision of "exclusion" and reduced the probability of the c-cost decision of "no decision" or "either". Thus, we should increase $n_A^\star$ to be at least as large as $n_0^\star$. Likewise, if we again fix $n_0^\star$ and increase $n_A^\star$ under H_A we increase the probability of "exclusion" at the expense of the probability of "either", both of which have cost c so the overall risk given H_A is unaffected. Similar reasoning can be used in the scenario illustrated in Fig. 4(b) to see that $n_0^\star$ must be at least as large as $n_A^\star$ to minimize the risk. Thus, under the loss function in Table 1(b) the Bayes risk is minimized for $n_0^\star = n_A^\star$, for any value of π.

4.3 Decision Analysis for Intervals and Limits

In Section 4.2 we illustrated how decision theoretic analysis can be used to derive a detection criterion. It is important to emphasize that this construction does not aim to control the probability of a false detection, as in Equation 5. Instead the goal is to control the overall expected loss of the procedure. Of course, if we specify $C \gg c$, false detections will be far less frequent than false negatives. Because we can always construct a confidence interval by inverting a test (as the set of values of λ_0 such that we cannot reject $H_0 : \lambda_S = \lambda_0$), the decision theoretic framework for detection leads to a confidence interval for the source intensity. The coverage of an interval derived from inverting a test is a function of the test's probability of a false positive: if the probability of a false positive is less than α the coverage of the resulting interval will be greater than $1 - \alpha$. Since the decision theoretic approach does not aim to control the probability of a false positive, however, the coverage of the resulting interval will vary.

A better strategy is to specify a loss function to directly quantify the desired properties of the interval or limit. For example, for a interval we might use

$$\text{Loss} = b \times \text{length}(\text{interval}) - I\{\text{interval contains } \theta\}$$

and for an upper limit we might use

$$\text{Loss} = b \times \text{limit} - I\{\theta < \text{limit}\},$$

where θ is a generic parameter of interest, $I\{\text{condition}\}$ is one if the condition is true and is zero otherwise, and b is a tuning parameter that specifies the relative importance of length and coverage. Let $[L(Y), U(Y)]$ be a generic interval computed from data Y. The risk of the interval can be written

$$\text{Risk}(\theta) = b \times \Big\{\text{E}(U(Y)|\theta) - \text{E}(L(Y)|\theta)\Big\} - \Pr\Big\{\theta \in [L(Y), U(Y)] \mid \theta\Big\},$$

where the second term on the right is the coverage. Notice that if we take b equal to zero the risk depends only on the coverage and the optimal interval is the entire parameter space (e.g., $(-\infty, +\infty)$). If we take b equal to ∞, the risk only depends on expected length and the optimal interval has $L(Y) = U(Y)$. Both the expected length and the coverage may depend on the value of θ. The Bayes risk computes the average of both quantities using a distribution on θ.[2] The goal is then to find functions L and U that minimize the Bayes risk. This is generally accomplished by parameterizing L and U. For example in a symmetric problem, we might consider intervals of the form $\hat{\theta} \pm e\hat{\sigma}$, where $\hat{\theta}$ and $\hat{\sigma}$ are estimates of θ and its error. This reduces minimization of the Bayes risk to a one dimensional minimization over e.

[2] Frequentist decision theoretic procedures are available that avoid the use of a distribution on θ by deriving the maximum risk over all values of θ. The interval that minimizes this maximum risk is considered optimal in the *minimax* sense.

5 Summary

The most important aspect of any statistical analysis is the specification of an adequate model. The choice of the specific procedure and/or the choice of statistical paradigm (i.e., frequency-based, Bayesian, or other) are typically far less critical to the properties of the procedure and the ultimate outcome of the analysis. Thus, when a statistical analysis exhibits odd behavior, the first remedy must be model diagnostics, validation, and improvement rather than questioning the choice of statistical procedure under the apparently inadequate model. Decision theoretic analysis allows us to directly specify the statistical properties that we hope for in a procedure and the relative importance that we place on these properties. This strategy is ideally suited to deriving detection procedures, intervals, and limits that exhibit properties that are viewed as best facilitating progress on the ultimate scientific goals.

Acknowledgements. The author thanks Louis Lyons for many helpful conversations on statistical issues in high energy physics, the organizing committees of the *Workshop on Statistical Issues Relevant to Significance of Discovery Claims* held at the Banff International Research Station in July 2010 and of *Phystat 2011* held at CERN in January 2011 for invitations to participate in these stimulating meetings, and the US National Science Foundation for financial support of this work (DMS-09-07522).

References

[1] M. Mandelkern, Statistical Science **17**, 149 (2002).

[2] D. A. van Dyk, Statistical Science **17**, 164 (2002).

[3] L. Lyons, Annals of Applied Statistics **2**, 887 (2008).

[4] F. Garwood, Biometrika **28**, 437 (1936).

[5] V. L. Kashyap *et al.*, The Astrophysical Journal **719**, 900 (2010).

[6] L. Demortier, Power-Constrained Upper Limits to Solve the Sensitivity Problem, Unpublised Manuscript, 2010.

[7] J. Feldman, Gary and R. D. Cousins, Physical Review D **57**, 3873 (1998).

[8] A. L. Read, **81** (2000).

[9] L. Demortier, Open Issues in the Wake of Banff 2010, Presentation at Phystat 2011 (CERN, Geneva, Switzerland), 2011.

[10] L. Lyons, Comments on ‘Look Elsewhere Effect’, Unpublished manuscript., 2010.

[11] D. Cox, Discovery: A Statistical Perspective, Presentation at Phystat 2011 (CERN, Geneva, Switzerland), 2011.

[12] M. Woodroofe, Importance Sampling and Error Probabilities, Presentation at "Statistical Issues Relevant to Significance of Discovery Claims (BIRS, Banff, Alberta, Canada, `http://www.birs.ca/events/2010/5-day-workshops/10w5068`), 2010.

[13] L. Wasserman, Statistical Science **17**, 163 (2002).

[14] L. Lyons, Statistical Issues in Particle Physics Analysis, Presentation at "Statistical Issues Relevant to Significance of Discovery Claims (BIRS, Banff, Alberta, Canada, `http://www.birs.ca/events/2010/5-day-workshops/10w5068`), 2010.