The Astrophysical Journal, 719:900–914, 2010 August 10 doi:10.1088/0004-637X/719/1/900
© 2010. The American Astronomical Society. All rights reserved. Printed in the U.S.A.

ON COMPUTING UPPER LIMITS TO SOURCE INTENSITIES

Vinay L. Kashyap[1], David A. van Dyk[2], Alanna Connors[3], Peter E. Freeman[4], Aneta Siemiginowska[1], Jin Xu[2], and Andreas Zezas[5]

[1] Smithsonian Astrophysical Observatory, 60 Garden Street, Cambridge, MA 02138, USA; vkashyap@cfa.harvard.edu, asiemiginowska@cfa.harvard.edu
[2] Department of Statistics, University of California, Irvine, CA 92697-1250, USA; dvd@ics.uci.edu, jinx@ics.uci.edu
[3] Eureka Scientific, 2452 Delmer Street, Suite 100, Oakland, CA 94602-3017, USA; aconnors@eurekabayes.com
[4] Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA; pfreeman@cmu.edu
[5] Physics Department, University of Crete, P.O. Box 2208, GR-710 03, Heraklion, Crete, Greece; azezas@cfa.harvard.edu
Received 2010 February 4; accepted 2010 June 22; published 2010 July 23

ABSTRACT

A common problem in astrophysics is determining how bright a source could be and still not be detected in an observation. Despite the simplicity with which the problem can be stated, the solution involves complicated statistical issues that require careful analysis. In contrast to the more familiar confidence bound, this concept has never been formally analyzed, leading to a great variety of often ad hoc solutions. Here we formulate and describe the problem in a self-consistent manner. Detection significance is usually defined by the acceptable proportion of false positives (background fluctuations that are claimed as detections, or Type I error), and we invoke the complementary concept of false negatives (real sources that go undetected, or Type II error), based on the statistical power of a test, to compute an upper limit to the detectable source intensity. To determine the minimum intensity that a source must have for it to be detected, we first define a detection threshold and then compute the probabilities of detecting sources of various intensities at the given threshold. The intensity that corresponds to the specified Type II error probability defines that minimum intensity and is identified as the upper limit. Thus, an upper limit is a characteristic of the detection procedure rather than the strength of any particular source. It should not be confused with confidence intervals or other estimates of source intensity. This is particularly important given the large number of catalogs that are being generated from increasingly sensitive surveys. We discuss, with examples, the differences between these upper limits and confidence bounds. Both measures are useful quantities that should be reported in order to extract the most science from catalogs, though they answer different statistical questions: an upper bound describes an inference range on the source intensity, while an upper limit calibrates the detection process. We provide a recipe for computing upper limits that applies to all detection algorithms.

Key words: methods: data analysis – methods: statistical

1. INTRODUCTION

When a known or suspected source remains undetected at a prescribed statistical significance during an observation, it is customary to report the *upper limit* on its intensity. This limit is usually taken to mean the largest intrinsic intensity that a nominal source can have with an appreciable chance of going undetected. Or equivalently, it is the smallest intrinsic intensity the source could have before its detection probability falls below a certain threshold. We emphasize that the upper limit is not meant to be an estimate or even a bound on the intensity of the source, but rather it is a quantification of the power of the *detection procedure* to detect weak sources. Thus, it is a measure that characterizes the detection process. The endpoints of the confidence interval, by contrast, provide ranges of possible values for the source intensity rather than quantify the sensitivity of the procedure. While the concept of upper limits has generally been well understood by astronomers as a form of censoring (Isobe et al. 1986) with an intrinsic connection to the detectability of a source (Avni et al. 1980), there has not been a statistically meaningful description that encapsulates its reliance on detectability as well as statistical significance.

Moreover, while the term "upper limit" has been traditionally used in this manner, it has also been used in cases where a formal detection process is not applied (e.g., when a source is known to exist at a given location because of a detection made in some other wavelength). In such cases, the upper edge of the confidence interval is derived and noted as the upper limit, regardless of the detectability of that source. In order to prevent confusion, we shall henceforth refer to the upper edge of a confidence interval as the "*upper bound*." Despite the intrinsic differences, numerous studies have described the computation of the *upper bound* as a proxy for the *upper limit* in increasingly sophisticated ways. The parameter confidence interval is statistically well understood and is in common use. Kraft et al. (1991) and Marshall (1992), for example, applied Poisson Bayesian-likelihood analysis to X-ray counts data to determine the credible range and thus set an upper bound on the source intensity. Feldman & Cousins (1998) recognized that the classical confidence interval at a given significance is not unique, and devised a scheme to determine intervals by comparing the likelihoods of obtaining the observed number of counts with the maximum likelihood estimate of the intensity and a nominal intensity; this procedure produces unique intervals where the lower edge overlaps with zero when there are very few counts, and the upper edge stands as a proxy for an upper limit. Variations in the background were incorporated via a sophisticated Bayesian analysis by Weisskopf et al. (2007).

The similarity of nomenclature between upper limits and upper bounds has led to considerable confusion in the literature on the nature of upper limits, how to compute them, and what type of data to use to do so. Many techniques have been used

to determine upper limits. It is not feasible to list all of these,[6] but for the sake of definiteness, we list a few methods culled from the literature: the techniques range from using the root-mean-square deviations in the background to set the upper limit (Gilman et al. 1986; Ayres 1999, 2004; Perez-Torres et al. 2009), adopting the source detection threshold as the upper limit (Damiani et al. 1997; Erdeve et al. 2009; Rudnick & Lemmerman 2009), computing the flux required to change a fit statistic value by a significant amount (Loewenstein et al. 2009), computing the *p*-value for the significance of a putative detection in the presence of background (Pease et al. 2006; Carrera et al. 2007), and identifying the upper limit with the parameter confidence bound (Schlegel & Petre 1993; Hughes et al. 2007). Here, we seek to clarify these historically often used terms in a statistically rigorous way.

Our goal here is to illustrate the difference between upper *limits* and upper *bounds*, and to develop a self-consistent description for the former that can be used with all extant detection techniques. Bounds and limits describe answers to different statistical questions, and usually both should be reported in detection problems. We seek to clarify their respective usage here. We set out the requisite definitions and statistical foundations in Section 2. In Section 3, we discuss the critical role played by the detection threshold in the definition of an upper limit and compare upper limits with upper bounds of confidence intervals. In Sections 2 and 3, we use a simple Poisson detection problem as a running example to illustrate our methods. In Section 4, we apply them to a signal-to-noise detection problem. Finally, we summarize in Section 5.

2. STATISTICAL BACKGROUND

Here we begin by describing our notation, and then discuss the nuances of the familiar concepts of confidence intervals, hypothesis testing, and statistical power. A glossary of the notation used is given in Table 1.

2.1. Description of the Problem

Our study is carried out in the context of background-contaminated detection of point sources in photon counting detectors, as in X-ray astronomical data. We set up the problem for the case of uncomplicated source detection (i.e., ignoring source confusion, intrinsic background variations, and instrumental effects such as vignetting, detector efficiency, PSF structure, bad pixels, etc). However, the methodology we develop is sufficiently general to apply in complex situations.

There is an important, subtle, and often overlooked distinction between an upper limit and the upper bound of a confidence interval, and the primary goal of this paper is to illuminate this difference. The confidence interval is the result of inference on the source intensity, while the upper limit is a measure of the power of the detection process. We can precisely state this difference in the context of an example. Suppose that we have a typical case of a source detection problem, where counts are collected in a region containing a putative or possible source and are compared with counts from a source-free region that defines the background. If the source counts exceed the threshold for detection, the source is considered to be detected. This detection threshold is usually determined by limiting the probability of a false detection. If the threshold were lower, there would be more false detections. Given this setup, we might ask how bright must a source be in order to ensure detection. Although statistically there are no guarantees, the upper limit is the minimum brightness that ensures a certain *probability of detection*. Critically, this value can be computed before the source counts are observed. It is based on two probabilities, (1) the probability of a false detection which determines the detection threshold and (2) the minimum probability that a bright source is detected. Although the upper limit is primarily of interest when the observed counts are less than the detection threshold, it does not depend on the observed counts. This is in sharp contrast to a confidence interval for the source intensity, that is typically of the form "source intensity estimate plus or minus an error bar," where the estimate and error bars depend directly on the observed source counts in the putative source region. Of course, the functional form of the confidence interval may be more complicated than in this example, especially in low count settings, but any reasonable confidence interval depends on the source counts, unlike upper limits. The fact that upper limits do not depend on the source counts while confidence intervals do should not be viewed as an advantage of one quantity or the other. Rather it reflects their differing goals. Upper limits quantify the power of the detection procedure and confidence intervals describe likely values of the source intensity. These distinctions are highlighted with illustrative examples in Section 3.1.1.

To formalize discussion, suppose that a known source has an intrinsic intensity in a given passband of λ_S and that the background intensity under the source is λ_B. Further suppose that the source is observed for a duration τ_S and that n_S counts are collected, and similarly, a separate measurement of the background could be made over a duration τ_B and n_B counts are collected. If the background counts are collected in an area

Table 1
Symbols and Notation

Symbol	Description
n_S	Counts observed in source area
n_B	Counts observed in background area
λ_S	Source intensity
λ_B	Background intensity
Λ_B	Range in background intensity λ_B
τ_S	Exposure time
τ_B	Exposure time for the background
r	Ratio of background to source area
$\mathcal{S}$	Statistic for hypothesis test
$\mathcal{S}^\star$	Detection threshold value of statistic $\mathcal{S}$
$n_S^\star$	Detection threshold value of statistic n_S
$\mathcal{U}$	Upper limit
α	The maximum probability of false detection
β	The probability of a detection
$\beta_{\rm min}$	The minimum probability of detection of a source with $\lambda_S = \mathcal{U}$
Pr(.)	Probability of
$n \sim f(.)$	Denoting that n is sampled from the distribution $f(.)$
Poisson(λ)	Poisson distribution with intensity λ
$\mathcal{N}(\mu, \sigma)$	Gaussian (i.e., normal) distribution with mean μ and variance σ^2

[6] An ADS query on astronomy abstracts, within the past year (excluding arXiv), containing "upper limit," yields roughly two papers per day (759). A quick survey shows this term used in several disparate ways: some are upper bounds of confidence regions, often convolved with physics information to get the upper bound of a confidence region on (say) mass; others are clearly the theoretical power of a suggested test; yet others use "upper limits" from previous work to obtain, e.g., line slopes.

r times the source area,[7] we can relate the observed counts to the expected intensities,

$$n_B|(\lambda_B, r, \tau_B) \sim \text{Poisson}(r\tau_B\lambda_B)$$
$$n_S|(\lambda_S, \lambda_B, \tau_S) \sim \text{Poisson}\Big(\tau_S(\lambda_S + \lambda_B)\Big), \quad (1)$$

in the background and source regions respectively, where n_B and n_S are independent. For simplicity, we begin by assuming that λ_B is known.

2.2. Confidence and Credible Intervals

Confidence intervals give a set of values for the source intensity that are consistent with the observed data. They are typically part of the inference problem for the source intensity. The basic strategy is to compute an interval of parameter values so that on repeated observations a certain proportion of the intervals contain the true value of the parameter. It is in this "repeated-observation" sense that a classical confidence interval has a given probability of containing the true parameter. Bayesian credible intervals have a more direct probabilistic interpretation. They are computed by deriving the posterior probability distribution of the source intensity parameter given the observed counts and finding an interval with nominal probability of containing the true rate (see, e.g., Loredo 1992; van Dyk et al. 2001; Kashyap et al. 2008). In summary, confidence intervals are *frequentist* in nature meaning that they are interpreted in terms of repeated observations of the source. Credible intervals, on the other hand, are *Bayesian* in nature meaning that they represent an interval of a certain posterior (or other Bayesian) probability.

2.2.1. Confidence Intervals

From a frequentist point of view randomness stems only from data collection—it is the data, not the parameters that are random. Often we use a 95% interval, but intervals may be at any *level* and we more generally refer to an L% confidence interval. Thus, the proper interpretation of a given interval is

> *L% of experiments (i.e., observations) with intervals computed in this way will result in intervals that contain the true value of the source intensity.*

In frequentist terms, this means that in any given experiment one cannot know whether the true source intensity is contained in the interval but if the experiment is repeated a large number of times, about L% of the resulting intervals will contain the true value. Strictly speaking, the more colloquial understanding that there is an L% chance that the "source intensity is contained in the reported confidence interval," is incorrect.

Put another way, a confidence interval for the source intensity gives values of λ_S that are plausible given the observed counts. Suppose that $\lambda_B = 3$ and that for each value of λ_S, we construct an interval $\mathcal{I}(\lambda_S)$ of possible values of the source counts that has at least an L% chance: $\Pr(n_S \in \mathcal{I}(\lambda_S)|\lambda_S, \lambda_B, \tau_S) \geqslant L\%$. Once the source count is observed and assuming λ_B is known, a confidence interval can be constructed as the set of values of λ_S for which the observed count is contained in $\mathcal{I}(\lambda_S)$,

$$\{\lambda_S : n_S \in \mathcal{I}(\lambda_S)\}. \quad (2)$$

In repeated observations, at least L% of intervals computed in this way cover the true value of λ_S.

The frequency coverage of confidence intervals is illustrated in Figure 1, where the confidence intervals of Garwood (1936) for a Poisson mean are plotted as boxes of width equal to the interval for various cases of observed counts (see van Dyk's discussion in Mandelkern 2002). For given values of λ_S and λ_B, the probability of the possible values of the observed counts can be computed using Equation (1); for each of these possible values, there will be a different confidence interval. Thus, the confidence intervals themselves have their own probabilities, which are represented as the heights of the boxes in Figure 1. Because these are 95% confidence intervals, the cumulative heights of the boxes that contain the true value of λ_S in their horizontal range must be at least 0.95. It is common practice to only report confidence intervals for detected sources so that only intervals corresponding to n_S above some threshold are reported. Unfortunately, this upsets the probability that the interval contains λ_S upon repeated observations. Standard confidence intervals are designed to contain the true value of the parameter (say) 95% of time, i.e., in 95% of data sets. If some of the confidence intervals are taken away (i.e., are not reported because, e.g., the counts are too small), there is no reason to expect that 95% of those remaining will contain the true value of the parameter. This is because instead of summing over all the values of n_S to get a probability that exceeds 95%, we are summing over only those values that are greater than the detection threshold. This results in a form of *Eddington bias* and is discussed in detail in Section 3.4.

2.2.2. Credible Intervals

In a Bayesian setting, probability is used to quantify uncertainty in knowledge and in this regard parameters are typically viewed as random quantities. This distinction leads to a more intuitive interpretation of the credible interval. A credible interval at the L% level, for example, is any interval that contains the true value of the parameter L% of the time according to its posterior distribution. (See Park et al. 2008 for a discussion on interval selection.) Thus, from a Bayesian perspective, it is proper to say that there is an L% chance that the source intensity is contained in the reported credible interval. The corresponding credible intervals look similar to the confidence intervals in Figure 1, at least in high count scenarios.[8]

So far we have considered a very simple problem with only one unknown parameter, λ_S. The situation is more complicated if there are unknown nuisance parameters, such as λ_B. In this case, frequency based intervals typically are constructed using asymptotic arguments and/or by conditioning on ancillary statistics that yield a conditional sampling distribution that does not depend on the nuisance parameter. Identifying ancillary statistics can be a subtle task and the resulting intervals may not be unique. Bayesian intervals, can be constructed using a simple

[7] For clarity, we assume that the expected intensities are in units of counts per unit time and that the source and background counts are collected over pre-specified areas in an image. However, our analysis is not restricted to this scenario. The nominal background-to-source area ratio r could include differences in exposure duration and instrument effective area. Furthermore, the nominal exposure duration could also incorporate effective area, e.g., to have units [photons count^{-1} cm^2 s], which implies that the expected intensity λ_S will have units [photons s^{-1} cm^{-2}]. Regardless of the units of λ_S and λ_B, the likelihood is determined by the Poisson distribution on the expected and observed counts as in Equation (1).

[8] In small count scenarios, Bayesian credible intervals may not exhibit their nominal frequency coverage. They do however have the proper Bayesian posterior probability.

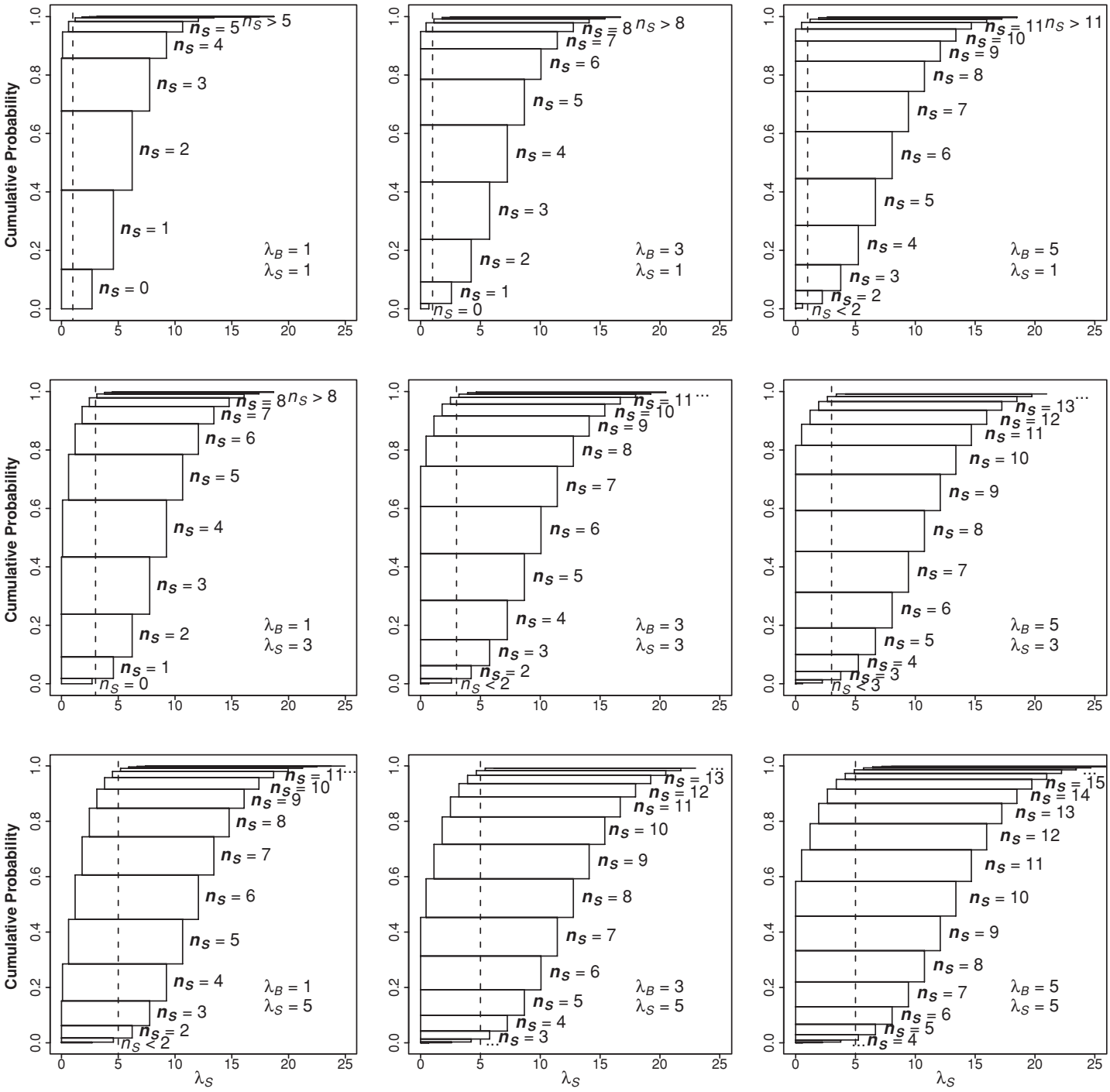

Figure 1. Confidence intervals for λ_S, computed for cases with different source and background intensities. The true value of λ_S is shown as a vertical dashed line and noted in the legend along with the true value of λ_B. We assume that λ_B is known exactly and adopt a nominal exposure $\tau_S = 1$ and background scaling $r = 10$. Each box corresponds to a different value of n_S. The horizontal width of each box denotes the width of the 95% confidence interval and the height denotes the probability of observing that many counts for a given λ_S. Top row: $\lambda_S = 1$ and for $\lambda_B = 1, 3, 5$ for the left, middle, and right columns, respectively. Middle row: as for the top row, for $\lambda_S = 3$. Bottom row: as for the top row, for $\lambda_S = 5$. The figure illustrates that if the models are correctly specified, very short intervals should be rare. Bayesian credible intervals (not shown) look similar, at least in high count scenarios.

and clear principle known as *marginalization*. For example, if λ_B is unknown, the marginal posterior distribution of λ_S is simply

$$p(\lambda_S|n_S, n_B, \tau_S, \tau_B, r) = \int p(\lambda_S, \lambda_B|n_S, n_B, \tau_S, \tau_B, r)d\lambda_B$$
$$= \int p(\lambda_S|\lambda_B, n_S, \tau_S)p(\lambda_B|n_B, \tau_B, r)d\lambda_B\,. \quad (3)$$

Credible intervals for λ_S are computed just as before, but using the marginal posterior distribution.[9]

[9] A popular frequentist alternative to the marginal posterior distribution is the profile likelihood function (see Park et al. 2008). Rather than averaging over nuisance parameters, the profile likelihood optimizes the likelihood over nuisance parameters for each value of the parameter of interest.

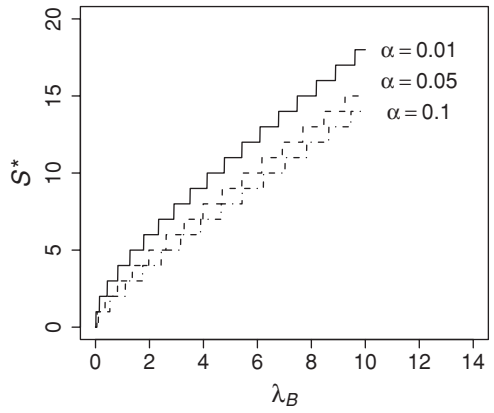

Figure 2. α-level detection threshold $\mathcal{S}^\star$ as a function of the background intensity λ_B, for the given α levels. Note that this is calculated assuming that the source intensity $\lambda_S = 0$. The detection threshold increases with increasing λ_B for a given α, and increases with decreasing α for a given λ_B.

2.3. Hypothesis Testing and Power

We emphasize that neither confidence nor credible intervals directly quantify the detection sensitivity of an experiment. To do this, we consider the detection problem in detail, which from a statistical point of view is a test of the hypothesis that there is no source emission in the given energy band,[10] i.e., a test of $\lambda_S = 0$. Formally, we test the *null hypothesis* that $\lambda_S = 0$ against the *alternative hypothesis* that $\lambda_S > 0$. The test is conducted using a test statistic that we denote by $\mathcal{S}$. An obvious choice for $\mathcal{S}$ is the counts in the source region, n_S; larger values of n_S are indicative of a detection of a source, since they become increasingly less likely to have been obtained as a random fluctuation from the background. Other choices for $\mathcal{S}$ are the signal-to-noise ratio (as in the case of sliding-cell local-detect algorithms; see Section 4) or the value of the correlation of a counts image with a basis function (as in the case of wavelet-based algorithms) or a suitably calibrated likelihood-ratio test statistic (as in the case of γ-ray detectors like *Fermi*; see Mattox et al. 1996). The count in the source region is an example of a test statistic that is stochastically increasing[11] in λ_S. For any fixed $\mathcal{S}^\star$, λ_B, τ_S, τ_B, and r, the probability that the test statistic $\mathcal{S}$ is less than the threshold $\mathcal{S}^\star$ decreases as λ_S increases, i.e., $\Pr(\mathcal{S} \leqslant \mathcal{S}^\star|\lambda_S, \lambda_B, \tau_S, r)$ decreases as λ_S increases. We assume that $\mathcal{S}$ is stochastically increasing in λ_S throughout.[12]

Because larger values of $\mathcal{S}$ indicate a source, we need to determine how large $\mathcal{S}$ must be before we can declare a source detection. This is done by limiting the probability of a false detection, also known as a Type I error. Thus, the detection threshold $\mathcal{S}^\star$ is the smallest value such that

$$\Pr(\mathcal{S} > \mathcal{S}^\star|\lambda_S = 0, \lambda_B, \tau_S, \tau_B, r) \leqslant \alpha, \tag{4}$$

where α is the maximum allowed probability of a false detection[13] and we declare a detection if the observed value of $\mathcal{S}$ is strictly greater than $\mathcal{S}^\star$:

> *If $\mathcal{S} \leqslant \mathcal{S}^\star$ we conclude there is insufficient evidence to declare a source detection.*
>
> *If $\mathcal{S} > \mathcal{S}^\star$ we conclude there is sufficient evidence to declare a source detection.*

We call $\mathcal{S}^\star$ the *α-level detection threshold* and sometimes write $\mathcal{S}^\star(\alpha)$ to emphasize its dependance on α (see Figure 2). Note that α is a bound on the probability of a Type I error; the actual probability of a Type I error is given by the probability on the left-hand side of Equation (4). Due to the discrete nature of the Poisson distribution, the bound is generally not achieved and the actual probability of a Type-I error is less than α.

Although its role in the definition of the detection threshold indicates that it is viewed as the more important concern, a *false detection*, also known as a "false positive," is not the only type of error. A *false negative*, or Type II error, occurs when a real source goes undetected (see Figure 3). The probability of a false negative is quantified through the power of the test to detect a source as a function of its intensity,

$$\beta(\lambda_S) = \Pr(\mathcal{S} > \mathcal{S}^\star|\lambda_S, \lambda_B, \tau_S, \tau_B, r)\,. \tag{5}$$

Equation (5) gives the probability of a detection. For any $\lambda_S > 0$, this is the power of the test or one minus the probability of a false

[10] There is a close relationship between confidence intervals and hypothesis testing. If the interval includes zero, this indicates that there is a real possibility of no source emission above the background and if the source has not been otherwise detected there may be no source at all. Conversely, in Appendix A we discuss how a hypothesis test can be inverted to construct a confidence interval.

[11] The term "stochastically increasing" means that there is a parameter (here λ_S) that defines a distribution of observable values (here n_S), and that all of the quantiles of n_S increase as λ_S increases. There is no guarantee that at any single instance of observation, a higher λ_S should lead to a higher n_S.

[12] In principle, the test statistic is only required to have different distributions under the alternative and null hypotheses. For simplicity, we assume that it tends to be larger under the alternative.

[13] In the simple Poisson counts case, the probability of a Type I error given in Equation (4) is computed as

$$\Pr(\mathcal{S} > \mathcal{S}^\star|\lambda_S = 0, \lambda_B, \tau_S, \tau_B, r) = \sum \left\{ \frac{e^{-\tau_S\lambda_B}(\tau_S\lambda_B)^{n'_S}}{n'_S!} \cdot \frac{e^{-r\tau_B\lambda_B}(r\tau_B\lambda_B)^{n'_B}}{n'_B!} \right\},$$

where the summation is over the set of values of (n'_S, n'_B) such that $\mathcal{S}(n'_S, n'_B) > \mathcal{S}^\star$ and we substitute $\lambda_S = 0$ into the mean of n_S given in Equation (1). Each term in the summation is a product of the likelihood of obtaining the specified counts in the absence of a source, given the background intensity and other observational parameters. In the simple case where $\mathcal{S}$ is the counts in the source region, n_S, and λ_B is known (i.e., n_B is not measured), this reduces to

$$\begin{aligned}\Pr(\mathcal{S} > \mathcal{S}^\star|\lambda_S = 0, \lambda_B, \tau_S) &= \sum_{n'_S=\mathcal{S}^\star+1}^{\infty} \frac{e^{-\tau_S\lambda_B}(\tau_S\lambda_B)^{n'_S}}{\Gamma(n'_S+1)} \\ &= 1 - \sum_{n'_S=0}^{\mathcal{S}^\star} \frac{e^{-\tau_S\lambda_B}(\tau_S\lambda_B)^{n'_S}}{\Gamma(n'_S+1)} \\ &= \frac{\gamma(\mathcal{S}^\star+1, \tau_S\lambda_B)}{\Gamma(\mathcal{S}^\star+1)},\end{aligned}$$

where

$$\gamma(\mathcal{S}^\star+1, \tau_S\lambda_B) = \Gamma(\mathcal{S}^\star+1)\int_0^{\tau_S\lambda_B} e^{-t}t^{n_S}dt$$

is the incomplete gamma function (see Equations (8.350.1) and (8.352.1) of Gradshteyn & Ryzhik 1980). In large count scenarios we may use continuous Gaussian distributions with their variances equal to their means in place of the discrete Poisson distributions in Equation (1); see Equation (13) in Section 4. In this case, we compute

$$\Pr(\mathcal{S}>\mathcal{S}^\star|\lambda_S = 0, \lambda_B, \tau_S, \tau_B, r) = \int\int \left\{ \frac{\exp\left[-\frac{(n'_S-\tau_S\lambda_B)^2}{\tau_S\lambda_B}\right]}{\sqrt{2\pi\tau_S\lambda_B}} \cdot \frac{\exp\left[-\frac{(n'_B-r\tau_B\lambda_B)^2}{r\tau_B\lambda_B}\right]}{\sqrt{2\pi r\tau_B\lambda_B}} \right\} dn'_S dn'_B,$$

where the integral is over the region of values of (n'_S, n'_B) such that $\mathcal{S}(n'_S, n'_B) > \mathcal{S}^\star$.

negative.[14] For $\lambda_S = 0$, Equation (5) gives the probability of a false detection (cf. Equation (4)) and consequently, $\beta(0) \leqslant \alpha$. This reflects the trade-off in any detection algorithm: the compromise between minimizing the number of false detections against maximizing the number of true detections. That is, if the detection threshold is set low enough to detect weaker sources, the algorithm will also produce a larger number of false positives that are actually background fluctuations. Conversely, the more stringent the criterion for detection, the smaller the probability of detecting a real source (this is illustrated by the location of the threshold $\mathcal{S}^\star$ that defines both α and β in Figure 3). Note that although our notation emphasizes the dependence of the power on λ_S, it also depends on λ_B, τ_S, τ_B, and r.

The power calculation is shown for the simple Poisson case in Figure 4, where $\beta(\lambda_S)$ is plotted for different instances of λ_B and for different levels of the detection threshold $\mathcal{S}^\star$. As expected, stronger sources are invariably detected. For a given source intensity, an increase in the background or a larger detection threshold (i.e., lower α) both cause the detection probability to decrease. In a typical observation, the background and the detection threshold are already known, and thus it is possible to state precisely the intensity λ_S at which the source will be detected at a certain probability. We may set a certain minimum probability, $\beta_{\rm min}$ of detecting a "bright" source by setting the exposure time long enough so that any source with intensity greater than a certain pre-specified cutoff has probability $\beta_{\rm min}$ or more of being detected. Conversely, we can determine how bright a source must be in order to have probability $\beta_{\rm min}$ or more of being detected with a given exposure time. This allows us to define an upper limit on the source intensity by setting a minimum probability of detecting the source. This latter calculation is the topic of Section 3.1 and the basis of our definition of an upper limit.

Power calculations are generally used to determine the minimum exposure time required to ensure a minimum probability of source detection (see Appendix B). In Section 3, we use them to construct upper limits.

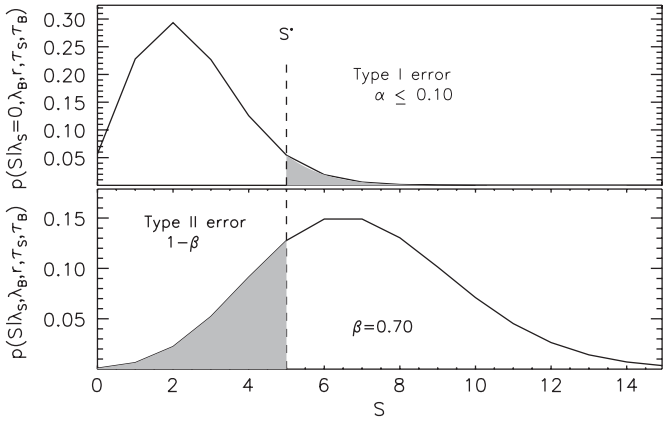

Figure 3. Illustration of Type I and Type II errors. A sketch of the probability distribution of test statistic $\mathcal{S}$ for specified values of the source and background intensities is shown for the simple case where $\mathcal{S} \equiv n_S$ and the background is known (see footnotes 13 and 14). The top panel depicts the probability $\Pr(\mathcal{S} = n_S|\lambda_S = 0, \lambda_B = 2, \tau_S = 1)$ and the bottom panel shows $\Pr(\mathcal{S} = n_S|\lambda_S = 5, \lambda_B = 2, \tau_S = 1)$. The vertical dashed line is a nominal detection threshold $\mathcal{S}^\star$ that corresponds to a significance of $\alpha \leqslant 0.1$, i.e., $\mathcal{S}^\star = 5$. The Type I error, or the probability of a false positive, is shown by the shaded region to the right of the threshold. (The actual Type I error for the adopted parameters is 0.05; values of $\mathcal{S}^\star$ less than 5 will cause the Type I error to exceed the specified significance.) The Type II error is the probability of a false negative and is shown (for $\lambda_S = 5$) by the shaded region to the left of the threshold in the lower panel. The detection probability of a source with intensity $\lambda_S = 5$ is $\beta = 0.7$ for this choice of α, for the given background intensity, and for the exposure time.

3. UPPER LIMITS

In this section, we develop a clear statistical definition of an upper limit that (1) is based on well-defined principles, (2) depends only on the method of detection, (3) does not depend on prior or outside knowledge about the source intensity, (4) corresponds to precise probability statements, and (5) is internally self-consistent in that all values of the intensity below the upper limit are less likely to be detected at the specified Type I error rate and values above are more likely to be detected.

[14] Here we use β to represent the power of the test, or one minus the probability of a Type II error. The statistical literature uses the notation β to denote either the Type II error (i.e., accepting the null hypothesis when it is false; e.g., Eadie et al. 1971) or for the power of the test itself (as we have done here; e.g., Casella & Berger 2002). As in the case of calculating α (see footnote 13), we can calculate β as

$$\Pr(\mathcal{S} > \mathcal{S}^\star|\lambda_S, \lambda_B, \tau_S, \tau_B, r) = \sum \left\{ \frac{e^{-\tau_S(\lambda_S+\lambda_B)}(\tau_S(\lambda_S+\lambda_B))^{n'_S}}{n'_S!} \cdot \frac{e^{-r\tau_B\lambda_B}(r\tau_B\lambda_B)^{n'_B}}{n'_B!} \right\},$$

where again the summation is over the set of values of (n'_S, n'_B) such that $\mathcal{S}(n'_S, n'_B) > \mathcal{S}^\star$. In the simple case where $\mathcal{S} = n_S$, λ_B is known, and n_B is not measured, we find

$$\Pr(\mathcal{S} > \mathcal{S}^\star|\lambda_S, \lambda_B, \tau_S) = \frac{\gamma(\mathcal{S}^\star + 1, \tau_S(\lambda_S + \lambda_B))}{\Gamma(n_S + 1)}.$$

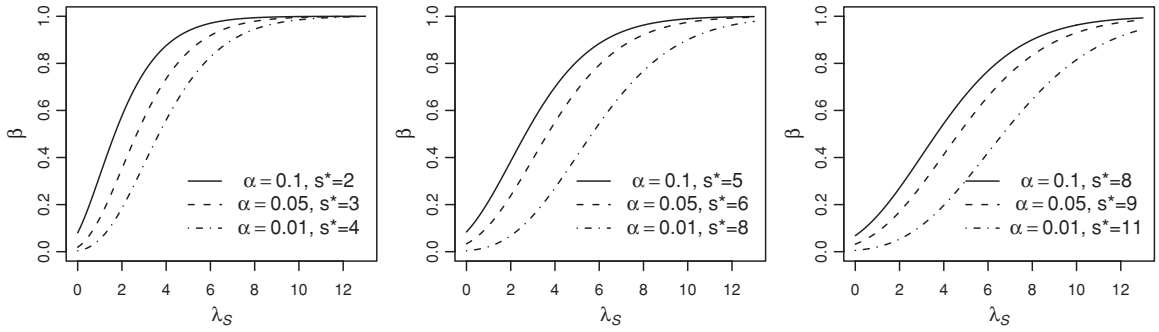

Figure 4. Power of the test, β, to detect a source as a function of the source intensity, λ_S, and detection threshold, $\mathcal{S}^\star$. The curves are calculated for different values of the background intensity (same values as in Figure 1), $\lambda_B = 1$ (left), $\lambda_B = 3$ (middle), and $\lambda_B = 5$ (right). The individual curves show $\beta(\lambda_S)$ for different $\mathcal{S}^\star$, each of which corresponds to a different bound on the probability of a Type I error, α, see Figure 2. The solid, dashed, and dash-dotted lines correspond to increasing detection thresholds, and decreasing values of α. As one would expect, β is higher for larger λ_S and lower λ_B, i.e., if the source is stronger or the background is weaker, it is easier to detect.

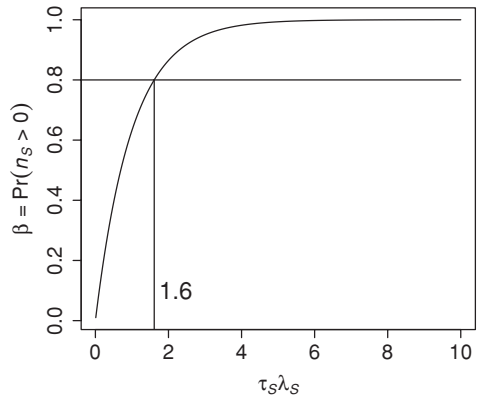

Figure 5. Upper limit with no background contamination. The figure plots $\Pr(n_S > 0|\lambda_S, \lambda_B = 0, \tau_S)$ as a function of $\tau_S\lambda_S$, thus giving $\mathcal{U}(\alpha, \beta_{\rm min})$ for any given τ_S and every value of $\beta_{\rm min}$. For example, reading across the line plotted at $\beta = 0.8$, gives $\tau_S\mathcal{U}(\alpha = 0.05, \beta_{\rm min} = 0.8) = 1.6$, which can be solved for the upper limit for any value of τ_S. Note that the upper limit decreases in inverse proportion to τ_S.

3.1. Definition

In astronomy, upper limits are inextricably bound to source detection: by an upper limit, an astronomer means

> *The maximum intensity that a source can have without having at least a probability of $\beta_{\rm min}$ 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 $\beta_{\rm min}$ of being detected under an α-level detection threshold.*

Unlike a confidence interval, the upper limit depends directly on the detection process and in particular on the maximum probability of a false detection and the minimum power of the test, that is on α and $\beta_{\rm min}$, respectively. In this way, an upper limit incorporates both the probabilities of a Type I and a Type II error. Formally, we define the upper limit, $\mathcal{U}(\alpha, \beta_{\rm min})$ to be the smallest λ_S such that

$$\Pr(\mathcal{S} > \mathcal{S}^\star(\alpha)|\lambda_S, \lambda_B, \tau_S, \tau_B, r) \geqslant \beta_{\rm min}. \quad (6)$$

Commonly used values for $\beta_{\rm min}$ throughout statistics are 0.8 and 0.9. If $\beta_{\rm min} \approx 1$, $\mathcal{U}(\alpha, \beta_{\rm min})$ represents the intensity of a source that is unlikely to go undetected, and we can conclude that an undetected source is unlikely to have intensity greater than $\mathcal{U}(\alpha, \beta_{\rm min})$.

The simplest example occurs in the hypothetical situation when λ_B is known to be zero and there is no background observation. In this case, we set $\mathcal{S} = n_S$ and note that $\Pr(n_S > 0|\lambda_S = 0, \lambda_B = 0, \tau_S) = 0$ so the detection threshold is zero counts and we declare a detection if there is even a single count. (Recall, we declare a detection only if $\mathcal{S}$ is strictly greater than $\mathcal{S}^\star$.) The upper limit in this case is the smallest value of λ_S with probability of detection greater than $\beta_{\rm min}$. Figure 5 plots $\Pr(n_S > 0|\lambda_S, \lambda_B = 0, \tau_S)$ as a function of $\tau_S\lambda_S$, thus giving $\mathcal{U}(\alpha, \beta_{\rm min})$ for any given τ_S and every value of $\beta_{\rm min}$. Note that the upper limit decreases in inverse proportion to τ_S.

When λ_B is greater than zero but well determined and can be considered to be known, the detection threshold using $\mathcal{S} = n_S$ is given in Equation (4). With this threshold in hand we can determine the maximum intensity a source can have with significant probability of not producing a large enough fluctuation above the background for detection. This is the upper limit.

In particular, $\mathcal{U}(\alpha, \beta_{\rm min})$ is the largest value of λ_S such that $\Pr(n_S \leqslant \mathcal{S}^\star(\alpha)|\lambda_S, \lambda_B, \tau_S) > 1 - \beta_{\rm min}$. This is illustrated for three different values of $\beta_{\rm min}$ (panels) and three different values of α (line types) in Figure 6. Note that the upper limit increases as $\beta_{\rm min}$ increases and as α decreases.

3.1.1. Illustrative Examples

To illustrate the difference between confidence bounds and upper limits, in the context of a detection process, we consider two simple examples. The first is an extreme case where the background intensity is known to be identically zero, and even one count in the source region would be classified as a detection. In this case, the upper limit is the smallest source intensity that can produce one count at a specified probability, e.g., a source with intensity of 5 generates one or more counts at a probability of ≈99.7% (see Section 3.1). In contrast, if one count is seen in the source region, the upper bound of an equal-tail 99.7% interval on the source intensity is 8.9 (Gehrels 1986). Thus, while similar in magnitude, it can be seen that upper bounds and upper limits are different quantities, describing different concepts.

Second, consider a more realistic case where the background is measured in a large region thought to be free of sources and scaled to the area covered by the source. Suppose that 800 counts are observed in an area 400 times larger than the source area and 3 counts are seen in the putative source region itself. The credible interval for the source intensity may be calculated at various significance levels (Section 2.2; see also van Dyk et al. 2001), and for this case we find that 68% credible interval with the lower bound at 0 is [0, 2.1], and the 99.7% interval is [0, 8.3]. But the question then arises as to whether the counts seen in the source region are consistent with a fluctuation of the observed background or not. Since at a minimum 7 counts are needed for a detection at a probability of 0.997 (corresponding to a Gaussian-equivalent "3σ" detection), it is considered that the source is not detected. The question then becomes how bright the source would have to be in order to be detected with a certain probability. Since a source of intensity 5.7 would have a 50% probability of producing sufficient counts for a detection, this sets an upper limit $\mathcal{U}(\alpha = 0.003, \beta_{\rm min} = 0.5) = 5.7$ counts on the undetected source's intensity (for a Type I error $\alpha = 0.003$ and a Type II error $\beta = 0.5$; see Section 3.1). Note that this limit is the same regardless of how many counts are actually seen within the source region, as expected from a quantity that calibrates the detection process. In contrast, the inference on the source intensity is always dependent on the number of observed source counts.

3.2. Unknown Background Intensity

So far our definition of an upper limit assumes that there are no unknown nuisance parameters, and in particular that λ_B is known. Unfortunately, the probabilities in Equations (4) and (6) cannot be computed if λ_B is unknown. In this section, we describe several strategies that can be used in the more realistic situation when λ_B is not known precisely.

The most conservative procedure ensures that the detection probability of the upper limit is greater than $\beta_{\rm min}$ for any possible value of λ_B. Generally speaking, the larger λ_B is, the larger λ_S must be in order to be detected with a given probability, and thus the larger the upper limit. Thus, a useful upper limit requires a finite range, Λ_B, to be specified for λ_B. Given this range, a

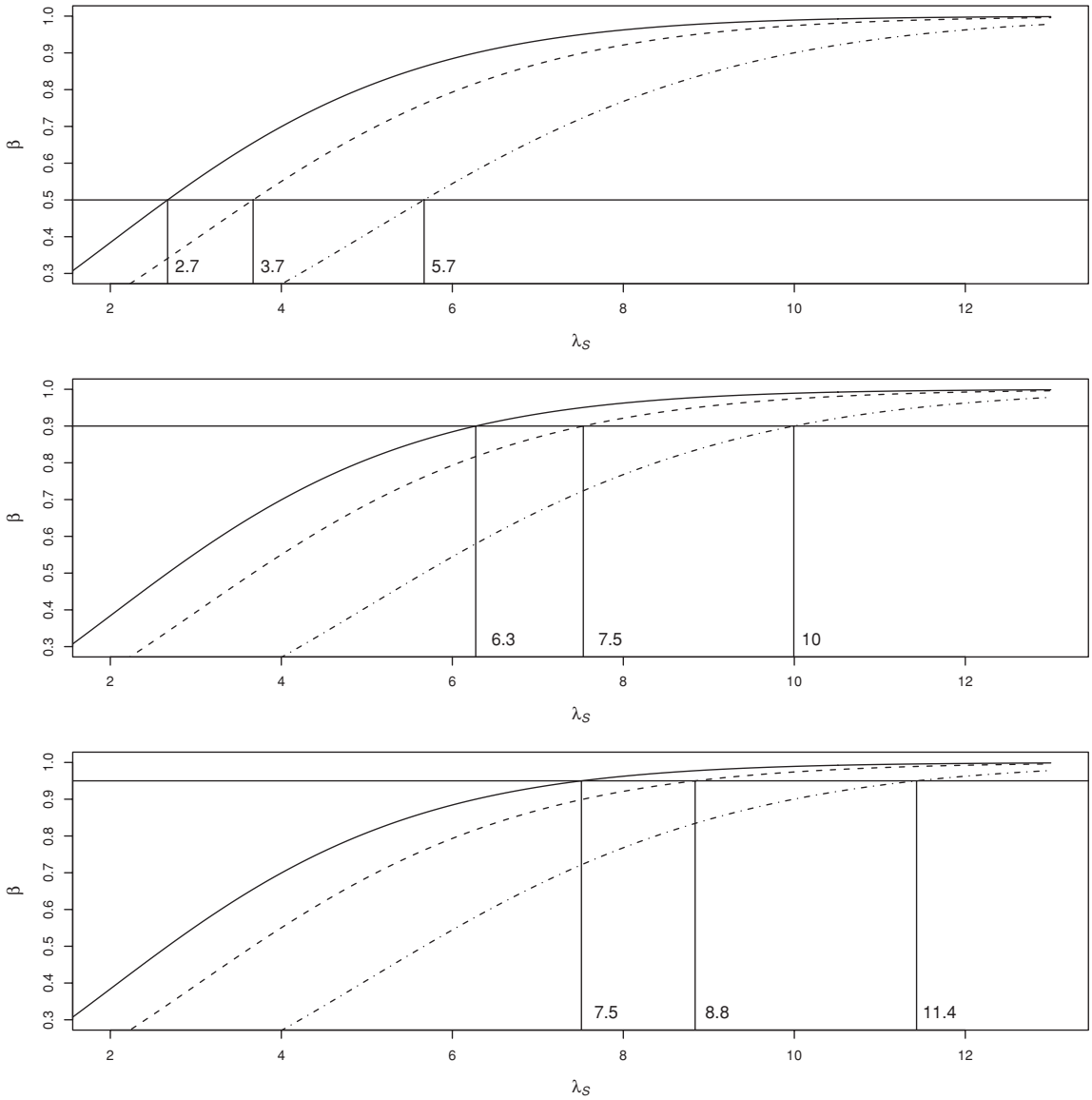

Figure 6. Computing upper limits based on the probability of detecting a source. The figure illustrates how upper limits may be defined for different probabilities of source detection under a given detection threshold. The curves correspond to $\beta(\lambda_S)$ for different values of $\mathcal{S}^\star$ and α: 5 and 0.1 (solid), 6 and 0.05 (dashed), and 8 and 0.01 (dash-dotted), and were all computed with $\lambda_B = 3$, as in the middle panel of Figure 4. Upper limits are computed by first adopting an acceptable probability for a source detection and then computing the intercept on λ_S of the $\beta(\lambda_S)$ curves. The panels show the value of the upper limits for the different values of $\mathcal{S}^\star$ for $\beta_{\rm min} = 0.5$ (top), $\beta_{\rm min} = 0.9$ (middle), and $\beta_{\rm min} = 0.95$ (bottom).

conservative upper limit can be defined as the smallest λ_S that satisfies

$$\inf_{\lambda_B \in \Lambda_B} \Pr(\mathcal{S} > \mathcal{S}^\star(\alpha) | \lambda_S, \lambda_B, \tau_S, \tau_B, r) \geqslant \beta_{\rm min}. \quad (7)$$

(We use the term *infimum* (inf) rather than *minimum* to allow for the case when the minimum may be on the boundary of, but outside, the range of interest. It is the largest number that is smaller than all the numbers in the range. For instance, the minimum of the range $\{x > 0\}$ is undefined, but the infimum is 0.) Unfortunately, unless the range of values Λ_B is relatively precise, this upper limit will often be too large to be useful.

In practice, there is a better solution. The background count provides information on the likely values of λ_B that should be used when computing the upper limit. In particular, the distribution of λ_B given n_B can be computed using standard Bayesian procedures[15] and used to evaluate the expected detection probability,

$$\beta(\lambda_S) = \int \Pr(\mathcal{S} > \mathcal{S}^\star(\alpha) | \lambda_S, \lambda_B, \tau_S, \tau_B, r) p(\lambda_B | n_B, \tau_B, r) d\lambda_B, \quad (8)$$

where $\mathcal{S}^\star(\alpha)$ is the smallest value such that

$$\int \Pr(\mathcal{S} > \mathcal{S}^\star(\alpha) | \lambda_S = 0, \lambda_B, \tau_S, \tau_B, r) p(\lambda_B | n_B, \tau_B, r) d\lambda_B \leqslant \alpha. \quad (9)$$

[15] In the presence of a nuisance parameter, frequentist procedures are more involved and typically require conditioning on an ancillary statistic, see, e.g., Appendix C. In the Bayesian case, the posterior distribution, $p(\lambda_B | n_B, \tau_B, r) \propto p(\lambda_B) p(n_B | \lambda_B, \tau_B, r)$, is the product of a prior distribution and the likelihood, normalized so that the posterior distribution integrates to 1. There are many choices of prior distributions available for λ_B, ranging from uniform on λ_B, to γ, to uniform in $\log(\lambda_B)$ (see, e.g., van Dyk et al. 2001).

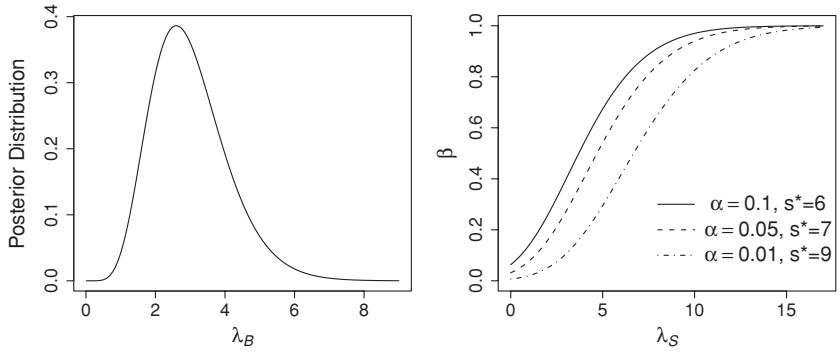

Figure 7. Upper limit with unknown background intensity. The left panel plots a Bayesian posterior distribution for λ_B, $p(\lambda_B|n_B, \tau_B, r) \propto \lambda_B^{6.5} e^{-\lambda_B/2.5}$, that is used in Equations (8) and (9) to compute β, the expected detection probability, as a function of λ_S with $\tau_S = \tau_B = r = 1$. The right panel plots $\beta(\lambda_S)$ for three values of α and their corresponding detection thresholds.

The upper limit is then computed as the smallest λ_S that satisfies $\beta(\lambda_S) \geqslant \beta_{\rm min}$. Unlike the upper limit described in Section 3.1, these calculations require data, in particular, n_B. For this reason, we call the smallest λ_S that satisfies Equation (8) the background count conditional upper limit or bcc upper limit. An intermediate approach that is more practical than using Equation (7) but more conservative than using Equation (8) is to simply compute a high percentile of $p(\lambda_B|n_B)$, perhaps its 95th percentile. The procedure for known λ_B can then be used with this percentile treated as the known value of λ_B. This is a conservative strategy in that it assumes a nearly worst case scenario for the level of background contamination.

As an illustration, suppose the uncertainty in λ_B given the observed background counts can be summarized in the posterior distribution plotted in the left panel of Figure 7. This is a gamma posterior distribution of the sort that typically arises when data are sampled from a Poisson distribution. Using this distribution, we can compute $\mathcal{S}^\star$ for any given value of α as the smallest value that satisfies Equation (9); the results are given for three values of α in the legend of Figure 7. We then use Equation (8) to compute $\beta(\lambda_S)$ as plotted in the right panel of Figure 7. The upper limit can be computed for any $\beta_{\rm min}$ using these curves just as in Figure 6.

3.3. Confidence Intervals versus Upper Limits

Although the form of a confidence interval makes it tempting to use its upper bound in place of an upper limit, this is misleading and blurs the distinction between the power of the detection procedure and the confidence with which the flux is measured. As an illustration, we have computed the upper limit for each of the nine panels in Figure 1 (using the true values of λ_B reported in each panel). The results are plotted as solid vertical lines in Figure 8. Note that unlike the upper bound of the confidence interval the upper limit does not depend on n_S and takes on values that only depend on λ_B. Using the upper bound of the confidence interval sometimes overestimates and sometimes underestimates the upper limit. In all but one of the nine cases with the highest λ_S/λ_B (i.e., $\lambda_S = 5$, $\lambda_B = 1$) the upper limit for λ_S is larger than λ_S. Of the nine cases, this is the one that is mostly likely to result is a source detection and is the only one with a probability of detection greater than $\beta_{\rm min} = 0.8$.

Alternatively, we can compute the value of $\beta_{\rm min}$ required for the upper bound of Garwood's confidence interval to be interpreted as an upper limit. Figure 9 does this for the three values of λ_B used in the three columns of Figures 1 and 8. Consider how the upper bound of Garwood's confidence interval increases with n_S in Figure 8. Each of these upper bounds can be interpreted as an upper limit, but with an increasing minimum probability of a source detection, $\beta_{\rm min}$. The three panels of Figure 8 plot how the required $\beta_{\rm min}$ increases with n_S for three values of λ_B. Note that a source with intensity equal to the upper bound of Garwood's confidence interval can have a detection probability as low as 20% or essentially as high as 100%. Thus, the upper bound does not calibrate the maximum intensity that a source can have with appreciable probability of going undetected in any meaningful way.

3.4. Statistical Selection Bias

As mentioned in Section 2.2.1, it is common practice to only report a confidence interval for detected sources. Selectively deciding when to report a confidence interval in this way can dramatically bias the coverage probability of the reported confidence interval.[16] We call this bias a *statistical selection bias*. Note that this is similar to the Eddington bias (Eddington 1913) that occurs when intensities are measured for sources close to the detection threshold. For sources whose intrinsic intensity is exactly equal to the detection threshold, the average of the intensities of the detections will be overestimated because downward statistical fluctuations result in non-detections and thus no intensity measurements. In extreme cases, this selection bias can lead to a nominal 95% confidence interval having a true coverage rate of well below 25%, meaning that only a small percentage of intervals computed in this way actually contain λ_S. As an illustration, Figure 10 plots the actual coverage of the nominal 95% intervals of Garwood (1936) for a Poisson mean when the confidence intervals are only reported if a source is detected with $\alpha = 0.05$. These intervals are derived under the assumption that they will be reported regardless of the observed value of n_S. Although alternative intervals could in principle be derived to have proper coverage when only reported for detected sources, judging from Figure 10 such intervals would have to be wider than the intervals plotted in Figure 1. It is critical that if standard confidence intervals are reported they be reported regardless of the observed value of n_S and regardless of whether a source is detected.[17]

[16] A similar concern was raised by Feldman & Cousins (1998) who noticed that deciding between a one-sided and a two-sided confidence interval can bias the coverage probability of the resulting interval, if the decision is based on the observed data.

[17] Though this is usually not feasible when sources are detected via an automated detection algorithm such as `celldetect` or `wavdetect`. However, in many cases, source detectability is determined based on a pre-existing catalog, and in such cases, both limits and bounds should be reported in order to not introduce biases into later analyses.

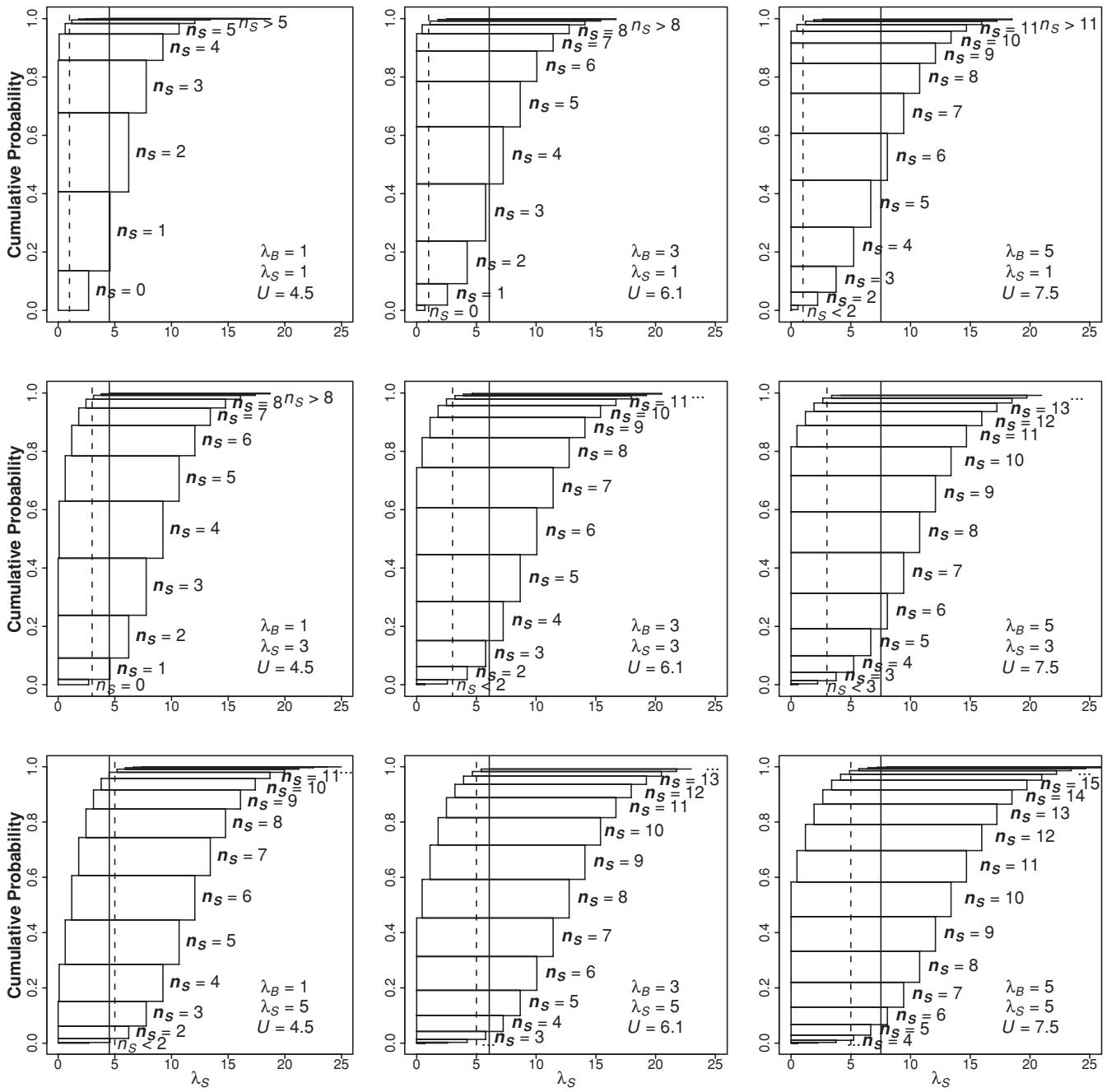

Figure 8. Illustrating the difference between confidence intervals and upper limits. This figure is identical to Figure 1, with an additional solid vertical horizontal line showing the location of the upper limit computed with $\beta_{\min} = 0.8$. The legend denotes the true value of λ_S, the assumed known value of λ_B, and the computed upper limit, $\mathcal{U}$. Note that unlike the confidence interval, which depends strongly on the number of observed source counts, n_S, the upper limit is fixed once the detection threshold $\mathcal{S}^\star$ (which depends on λ_B) and the minimum detection probability $\beta_{\min}$ are specified.

3.5. The Detection Threshold as an Upper Limit

As discussed in Section 1, the detection threshold is sometimes used as an upper limit. Under certain circumstances, this can be justified under our definition of an upper limit. Suppose that some invertible function $f(\mathcal{S})$ can be used as an estimate of λ_S and that for any λ_B, the sampling distribution of $f(\mathcal{S})$ is continuous with median equal to λ_S. That is,

$$\Pr(f(\mathcal{S}) > \lambda_S | \lambda_S, \lambda_B, \tau_S, \tau_B, r) = 0.5. \quad (10)$$

Because Equation (10) holds for any λ_S, it holds for $\lambda_S = f(\mathcal{S}^\star)$. That is,

$$\Pr(f(\mathcal{S}) > f(\mathcal{S}^\star) | \lambda_S = f(\mathcal{S}^\star), \lambda_B, \tau_S, \tau_B, r) = 0.5. \quad (11)$$

Inverting f and integrating over $p(\lambda_B | n_B, \tau_B, r)$, we have

$$\int \Pr(\mathcal{S} > \mathcal{S}^\star | \lambda_S = f(\mathcal{S}^\star), \lambda_B, \tau_S, \tau_B, r) p(\lambda_B | n_B, \tau_B, r) d\lambda_B = 0.5. \quad (12)$$

Comparing Equation (12) with Equation (8), we see that $f(\mathcal{S}^\star) = \mathcal{U}(\alpha, \beta_{\min} = 0.5)$. Thus, if f is an identity function the detection threshold is an upper limit. Although the assumption that the sampling distribution of $f(\mathcal{S})$ has median λ_S for every λ_B, is unrealistic in the Poisson case, it is quite reasonable with Gaussian statistics. Even if this assumption holds, $f(\mathcal{S}^\star)$ is a weak upper limit in that half the time a source with this intensity would go undetected and there is a significant chance

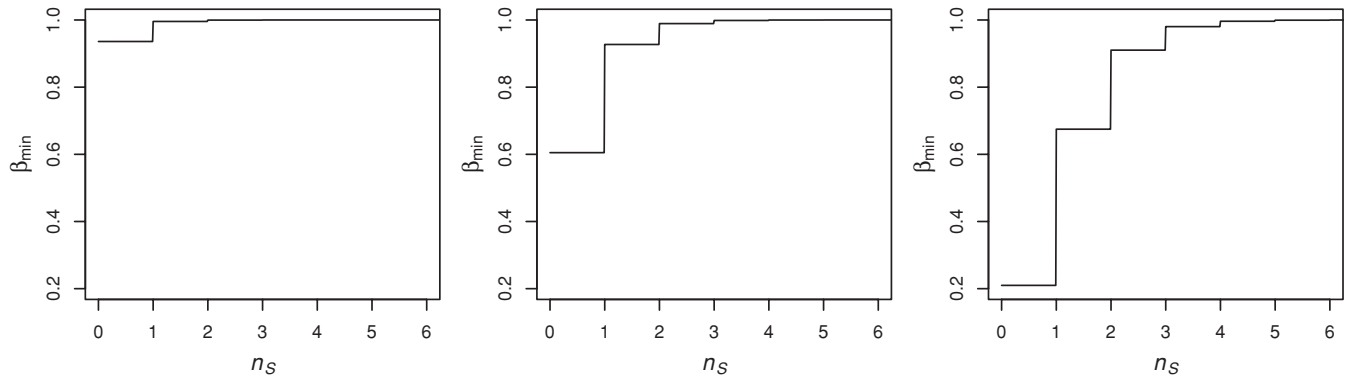

Figure 9. Interpreting upper bounds as upper limits. The three panels plot the probability or source detection for a source with λ_S equal to the upper bound of Garwood's confidence intervals. Because the confidence intervals depend on n_S, the detection probability, $\beta_{\rm min}$ increases with n_S. The three panels correspond to $\lambda_B = 1$ (left), $\lambda_B = 3$ (middle), and $\lambda_B = 5$ (right), as in the columns of Figure 8. All calculations were preformed with $\tau_S = \tau_B = r = 1$. Because a source with intensity equal to the upper bound can have a detection probability as low as 20% or as high as 100%, the upper bound does not calibrate the maximum intensity that a source can have with appreciable probability of going undetected in any meaningful way.

that sources with intrinsic intensity larger than $f(\mathcal{S}^\star)$ would remain undetected.

It should be emphasized that even when the detection threshold is used as an upper limit, it is not an "upper limit on the counts," but an upper limit on the intrinsic intensity of the source. The counts are an observed, not an unknown quantity. There is no need to compute upper bounds, error bars, or confidence intervals on known quantities. It is for the unknown source intensity that these measures of uncertainty are useful.

3.6. Recipe

Our analysis of upper limits and confidence interval assumes that the observables are photon counts that we model using the Poisson distribution. However, the machinery we have developed is applicable to any process that uses a significance-based detection threshold. Here, we briefly set out a general recipe to use in more complicated cases. For complex detection algorithms, some of the steps may require Monte Carlo methods.

1. Define a probability model for the observable source and background data set given the intrinsic source and background strengths, λ_S and λ_B, respectively. For the simple Poisson case, this is defined in Equation (1). In many applications, these could be approximated using Gaussian distributions. It is typically required that a background data set be observed but in some cases λ_B may be known a priori. The source could be a spectral line, or an extra model component in a spectrum, or possibly even more complex quantities that are not directly related to the intensity of a source.
2. Define a test statistic $\mathcal{S}$ for measuring the strength of the possible source signal. In the simple Poisson case, we set $\mathcal{S} = n_S$. The "source" could be a spectral line or any extra model component in a spectrum, or a more complex quantity that is not related to the intensity of a source.
3. Set the maximum probability of a false detection, α, and compute the corresponding α-level detection threshold, $\mathcal{S}^\star$. Although $\mathcal{S}^\star$ depends on λ_B, we can compute the expected $\mathcal{S}^\star$ by marginalizing over λ_B using $p(\lambda_B|n_B)$ when λ_B is not known exactly. Likewise, if λ_S is defined as a function of several parameters, the same marginalization procedure can be used to marginalize over any nuisance parameters. In this case, we typically marginalize over $p(\eta|n_B)$ or perhaps $p(\eta|n_S, n_B)$, where η is the set of nuisance parameters. In this regard, we are setting α to be a quantile of the posterior predictive distribution of $\mathcal{S}$, under the constraint that $\lambda_S = 0$; see Gelman et al. (1996) and Protassov et al. (2002).
4. Compute the probability of detection, $\beta(\lambda_S)$, for the adopted detection threshold $\mathcal{S}^\star$.
5. Define the minimum probability of detection at the upper limit, $\beta_{\rm min}$. Traditionally, $\beta_{\rm min} = 0.5$ has been used in conjunction with $\alpha = 0.003$ in astronomical analysis (see Section 3.5).
6. Compute the smallest value of λ_S such that $\beta(\lambda_S) \geqslant \beta_{\rm min}$. This is the *upper limit*.

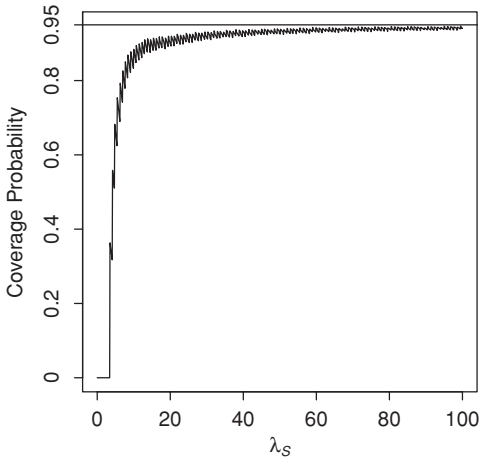

Figure 10. Conditional coverage probability of confidence interval reported only for detected sources. When confidence intervals are only reported for detected sources the coverage probability may be very different than the nominal level of the interval. This plots shows the true coverage of the 95% nominal intervals of Garwood (1936) when they are only reported for sources detected with $\alpha = 0.05$ (with $\lambda_B = 3$). For small values of λ_S the coverage can be very low and as λ_S grows, the coverage probability converges to 95%. Far fewer intervals contain λ_S than one would expect given the nominal level. Confidence intervals must be reported regardless of n_S and regardless of whether a source is detected. The jagged appearance of curve stems from the discrete nature of Poisson data.

4. EXAMPLE: SIGNAL-TO-NOISE RATIO

We focus below on signal-to-noise (S/N)-based detection at a single location as an example application. The S/N was the primary statistic used for detecting sources in high-energy astrophysics before the introduction of maximum-likelihood and wavelet-based methods. Typically, S/N = 3 was used as the detection threshold, corresponding to $\alpha = 0.003$ in the Gaussian regime. Here, we apply the recipe in Section 3.6 to derive an upper limit with S/N-based detection. Our methods can also be applied to more sophisticated detection algorithms such as sliding-cell detection methods such as `celldetect` (Harnden et al. 1984; Dobrzycki et al. 2000; Calderwood et al. 2001), and wavelet-based detection methods such as `pwdetect` (Damiani et al. 1997), `zhdetect` (Vikhlinin et al. 1997), `wavdetect` (Freeman et al. 2002), etc. Implementation of our technique for these methods will vary in detail, and we leave these developments for future work.

We begin with a Gaussian probability model for the source and background counts (Step 1 in Section 3.6)

$$\begin{aligned} n_B|(\lambda_B, r, \tau_B) &\sim \mathcal{N}(\mu = r\tau_B\lambda_B,\ \sigma = \sqrt{r\tau_B\lambda_B}) \\ n_S|(\lambda_S, \lambda_B, \tau_S) &\sim \mathcal{N}(\mu = \tau_S(\lambda_S + \lambda_B),\ \sigma = \sqrt{\tau_S(\lambda_S + \lambda_B)}), \end{aligned} \tag{13}$$

where λ_B and λ_S are non-negative. We assume that the source is entirely contained within the source cell and that the PSF does not overlap the background cell. We can estimate λ_B and λ_S by setting n_B and n_S to their expectations (method of moments), as

$$\hat{\lambda}_B = \frac{n_B}{r\tau_B} \text{ and } \hat{\lambda}_S = \frac{n_S}{\tau_S} - \frac{n_B}{r\tau_B}. \tag{14}$$

The variance of $\hat{\lambda}_S$ is

$$\mathrm{var}(\hat{\lambda}_S) = \frac{\lambda_S}{\tau_S} + \frac{(\tau_S + r\tau_B)\lambda_B}{r\tau_S\tau_B} \tag{15}$$

which we can estimate by plugging in $\hat{\lambda}_S$ and $\hat{\lambda}_B$ as

$$\widehat{\mathrm{var}}(\hat{\lambda}_S) = \frac{n_S}{\tau_S^2} + \frac{n_B}{r^2\tau_B^2}. \tag{16}$$

To use the S/N as a detection criterion, we define (Step 2 in Section 3.6)

$$\mathcal{S} = \frac{\hat{\lambda}_S}{\sqrt{\widehat{\mathrm{var}}(\hat{\lambda}_S)}} = \frac{r\tau_B n_S - \tau_S n_B}{\sqrt{r^2\tau_B^2 n_S + \tau_S^2 n_B}}. \tag{17}$$

Step 3 in Section 3.6 says that the maximum probability of a false detection should be set and $\mathcal{S}^\star$ computed accordingly. Instead we adopt the standard detection threshold, $\mathcal{S}^\star = 3$, used with the S/N and compute the corresponding α,

$$\alpha(\lambda_B) = \int_{\mathcal{R}} p(n'_B|\lambda_B, r, \tau_B)\, p(n'_S|\lambda_S = 0, \lambda_B, \tau_S)\, dn'_B dn'_S, \tag{18}$$

where $\mathcal{R}$ is the region where $\mathcal{S}(n'_S, n'_B) > \mathcal{S}^\star = 3$ (see footnote 13). For given values of λ_B and r, the integral in Equation (18) can be easily evaluated via Monte Carlo. Alternatively, we can compute α by marginalizing over λ_B if it is unknown,

$$\alpha = \int_0^\infty \alpha(\lambda_B)\, p(\lambda_B|n_B, r, \tau_B)\, d\lambda_B, \tag{19}$$

where n_B is the observed background count and $\alpha(\lambda_B)$ is computed in Equation (18). The probability of detection, $\beta(\lambda_S)$ is computed (Step 4 in Section 3.6) by evaluating the same integral as in Equation (18) except that λ_S is not set to zero. We can make the same substitution in Equation (19) if λ_B is unknown. With $\beta_{\min}$ in hand (Step 5 in Section 3.6), we can find the value of λ_S such that $\beta(\lambda_S) = \beta_{\min}$ (Step 6 in Section 3.6). This is the upper limit.

Figure 11 illustrates the use of Equation (18) to compute $\beta(\lambda_B)$ for several values of λ_B, with $r = \tau_S = \tau_B = 1$. The upper limit is computed as the value of λ_S such that $\beta(\lambda_S) = \beta_{\min}$. The three panels of Figure 11 report the resulting upper limits for $\beta_{\min} = 0.5$, 0.9, and 0.95, respectively.

5. SUMMARY

We have carefully considered the concept of upper limits in the context of undetected sources, and have developed a rigorous formalism to understand and express this concept. Despite its seeming simplicity, upper limits are not treated in a uniform fashion in astronomical literature, leading to considerable variations in meaning and value. We formally define an upper limit to the source intensity as the maximum intensity it can have without exceeding a specified detection threshold at a given probability. This is defined by the statistical power of the detection algorithm. This is equivalent to defining it as the largest source intensity that remains undetected at the specified probability, and is defined by the probability of Type II error. Thus, if the detection probability is computed for a variety of source intensities, the upper limit is then identified by determining the intercept of the required probability with this curve. Thus, an upper limit is dependent only on the detection criterion, which is generally a function only of the background, and independent of the source counts. This is different from the upper bound (i.e, the upper edge of a confidence interval), which is obtained when the probability distribution of the source intensity is computed given that some counts are observed in the putative source region. We distinguish between the upper bound of the confidence interval and the upper limit of source detectability. Unlike a confidence interval (or Bayesian credible interval), an upper limit is a function of the detection procedure alone and does not necessarily depend on the observed source counts.

The primary goals of this paper are to clearly define an upper limit, to sharpen the distinction between an upper limit and an upper bound, and to lay out a detailed procedure to compute the former for any detection process. In particular, we have shown how to compute upper limits for the simple Poisson case. We also provide a step-by-step procedure for deriving it when a simplified significance-based detection method is employed. To extract the most science from catalogs, we argue for using a consistent, statistically reasonable recipe of an upper limit being related to the statistical power of a test. In addition, we illustrate the peril of using an upper bound in place of an upper limit and of only reporting a frequentist confidence interval when a source is detected. Conversely, including confidence bounds, even for non-detections, is a way to avoid the Eddington bias and increase the scientific usefulness of large catalogs.

We also describe a general recipe for calculating an upper limit for any well-defined detection algorithm. Briefly, the detection threshold should be first defined based on an acceptable probability of a false detection (the α-level threshold), and an intensity that ensures that the source will be detected at a specifed probability (the β-level detection probability) should

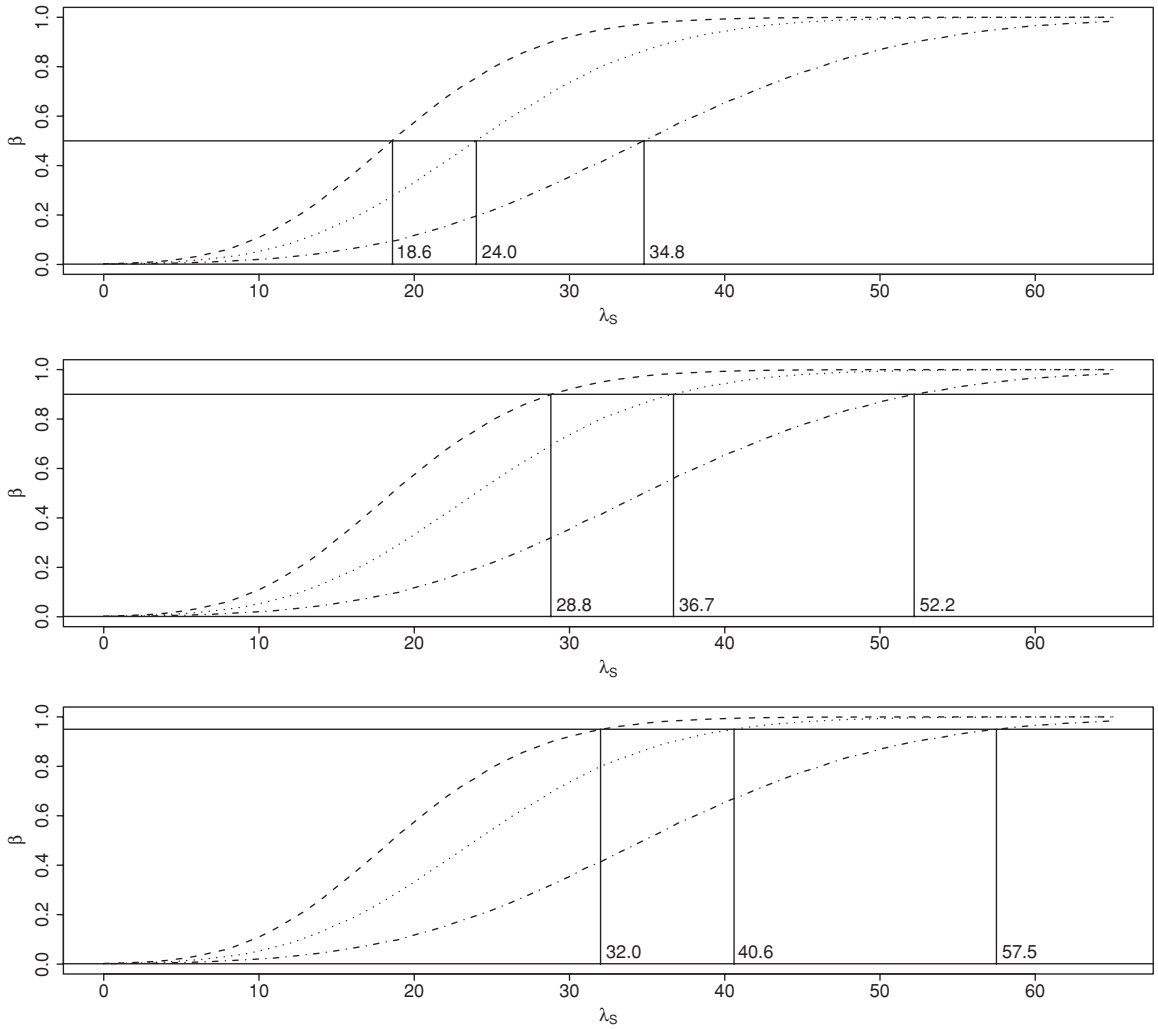

Figure 11. Computing upper limits based on the probability of S/N detection of a source. The curves in each panel correspond to the probability of source detection as a function of λ_S using an S/N detection threshold of $\mathcal{S}^\star = 3$. The curves were computed with $r = 1$, $\tau_S = 1$, $\tau_B = 1$, and with $\lambda_B = 10, 20$, and 50 (dashed, dotted, and dash-dotted lines, respectively). Upper limits are computed by first adopting an acceptable probability for a source detection, and then computing the intercept on λ_S of the $\beta(\lambda_S)$ curves. The panels show the value of the upper limits for the different values of λ_B for $\beta_{\rm min} = 0.5$ (top), $\beta_{\rm min} = 0.9$ (middle), and $\beta_{\rm min} = 0.95$ (bottom).

be computed; this latter intensity is identified with the upper limit. We recommend that when upper limits are reported in the literature, both the corresponding α and β values should also be reported.

This work was supported by NASA-AISRP grant NNG06G F17G (A.C.), CXC NASA contract NAS8-39073 (V.L.K., A.S.), NSF grants DMS 04-06085, and DMS 09-07522 (D.A.v.D.). We acknowledge useful discussions with Rick Harnden, Frank Primini, Jeff Scargle, Tom Loredo, Tom Aldcroft, Paul Green, Jeremy Drake, and participants and organizers of the SAMSI/ SaFeDe Program on Astrostatistcs.

APPENDIX A

CONSTRUCTING A CONFIDENCE INTERVAL BY INVERTING A HYPOTHESIS TEST

Here we discuss the relationship between confidence intervals and hypothesis tests and in particular how a hypothesis test can be used to construct a confidence interval.

A confidence interval reports the set of values of the parameter that are consistent with the data. When this set includes $\lambda_S = 0$ it means that the data are consistent with no source and we expect the null hypothesis not to be rejected and no source to be detected. There is a more formal relationship between confidence intervals and hypothesis testing and we can use a detection method to generate a confidence interval. Suppose that rather than testing the null hypothesis that $\lambda_S = 0$, we are interested in testing the more general null hypothesis that $\lambda_S \leqslant \lambda_S^\star$, where $\lambda_S^\star$ is any non-negative number. That is, we are interested in detecting only sources of at least a certain brightness. In this case, the detection threshold, $\mathcal{S}^\star(\lambda_S^\star)$, is defined as the smallest value such that

$$\Pr(\mathcal{S} > \mathcal{S}^\star(\lambda_S^\star)|\lambda_S = \lambda_S^\star, \lambda_B, \tau_S, \tau_B, r) \leqslant \alpha. \qquad \text{(A1)}$$

Given an observed value of $\mathcal{S}$, we can construct the set of values $\lambda_S^\star$ for which we cannot reject the null hypothesis that $\lambda_S \leqslant \lambda_S^\star$. This is a set of values of λ_S that are consistent with the data and they form a $100(1-\alpha)\%$ confidence interval. This particular, confidence interval, however is of the form $(a, +\infty)$: For any observed count there is a $\lambda_S^\star$ large enough so that we cannot reject the null hypothesis that $\lambda_S \leqslant \lambda_S^\star$. By reversing the null hypothesis to $\lambda_S \geqslant \lambda_S^\star$ we can obtain an interval of the form $(0, a)$ and by setting up a two-sided test of the null hypothesis

that $\lambda_S = \lambda_S^\star$ against the alternative hypothesis that $\lambda_S \neq \lambda_S^\star$ we can obtain an interval of the more common form (a, b).

APPENDIX B

THE RELATIONSHIP BETWEEN UPPER LIMITS AND THE POWER OF THE TEST

An upper limit turns around the usual use of the power of a test. Power is ordinarily used to determine the exposure time required to be sure that a source with intensity $\lambda_{S\min}$ or greater has at least probability $\beta_{\min}$ of being detected. That is, the smallest τ_S is found that satisfies Equation (6) for any $\lambda_S \geqslant \lambda_{S\min}$ and with $\lambda_{S\min}$ fixed in advance. Thus, power is used to design an observation so that we have at least a certain probability of detecting a source of given brightness. With an upper limit on the other hand τ_S is fixed and Equation (6) is solved for λ_S. This is illustrated in Figure 12 which plots τ_S versus λ_S with fixed λ_B, τ_B, r, α, and $\beta_{\min}$ and shows what values of τ_S and λ_S satisfy Equation (6) in the simple Poisson case. The shaded area above and to the right of the curve is where the detection probability exceeds $\beta_{\min} = 0.90$. Thus the curves give the upper limit (on the horizontal scale) as a function of the exposure time. The upper limit generally decreases as the exposure time τ_S increases, but not monotonically. Due to the discrete nature of Poisson data, the threshold value $\mathcal{S}^\star$ changes in integer steps to allow for the inequality in Equation (4) to be satisfied. This behavior may be graphically illustrated by considering the sketch of the relevant quantities in Figure 3. As τ_S increases, $\mathcal{S}^\star$ increases in steps, causing the probability of false detection to abruptly fall and then smoothly increase to α. As the expected background in the source region increases, the upper curve shifts to the right, thereby increasing the shaded area that lies above the threshold value. However, when the area of the shaded region in the upper plot becomes larger than the tolerable probability of a Type I error, $\mathcal{S}^\star$ must be increased by one to reduce that probability. As τ_S increases, the lower curve remains stationary while $\mathcal{S}^\star$ is unchanging. At this stage, the upper limit, $\mathcal{U}(\beta)$ is set as that value of λ_S which ensures that the Type II error is β (see Equation (6)), and thus slowly decreases as τ_S increases. When $\mathcal{S}^\star$ increases as a step function, the lower curve shifts to the right in order to maintain the same value of β, and the upper limit abruptly increases.

APPENDIX C

AN ALTERNATIVE METHOD FOR AN UNKNOWN BACKGROUND CONTAMINATION RATE

In the body of the article, we suggested conditioning on n_B in order to effectively estimate λ_B when it is unknown. A different strategy conditions instead on the total count $n_S + n_B$ in order to remove λ_B from the model. This method is based on the simple probabilistic result that if X and Y are independent Poisson variables with means λ_X and λ_Y, respectively, then given $X + Y$, the variable X follows a binomial distribution. Applying this result to n_S and n_B with Poisson models given in Equation (1), we have

$$n_S \mid (n_S + n_B, \lambda_S, \lambda_B, r, \tau_S, \tau_B) \sim \text{Binomial}\left(n_S + n_B, \frac{\tau_S(\lambda_S + \lambda_B)}{\tau_S\lambda_S + (\tau_S + r\tau_B)\lambda_B}\right), \quad \text{(C1)}$$

a binomial distribution with $n_S + n_B$ independent counts each with probability $\tau_S(\lambda_S + \lambda_B)/(\tau_S\lambda_S + (\tau_S + r\tau_B)\lambda_B)$ of being a source count. Reparameterizing (λ_S, λ_B) via $\xi_S\lambda_B = \lambda_S + \lambda_B$, Equation (C1) becomes

$$n_S \mid (n_S + n_B, \lambda_S, \lambda_B, r, \tau_S, \tau_B) \sim \text{Binomial}\left(n_S + n_B, \frac{\xi_S}{\xi_S + r\tau_B/\tau_S}\right), \quad \text{(C2)}$$

which does not depend on the unknown background intensity. Here $\xi_S = (\lambda_S + \lambda_B)/\lambda_B$, which is equal to 1 if there is no source and grows larger for brighter sources. Because Equation (C2) does not depend on λ_B it can be used for direct frequency based calculations even when λ_B is unknown. In particular, a detection threshold can be computed based on a test of the null hypothesis that $\xi_S = 1$, which is equivalent to $\lambda_S = 0$. This is done using Equation (4) with $\mathcal{S} = n_B$ and using distribution given in Equation (C2) with $\xi_S = 1$. In particular, we find the smallest $\mathcal{S}^\star$ such that $\Pr(n_S > \mathcal{S}^\star | n_S + n_B, \xi_S = 1, r, \tau_S, \tau_B) \leqslant \alpha$. With the detection threshold in hand, we can compute an upper limit for ξ_S using Equation (6). The upper limit is the smallest ξ_S such that $\Pr(n_S > \mathcal{S}^\star | n_S + n_B, \xi_S, r, \tau_S, \tau_B) \geqslant \beta_{\min}$. Unfortunately, this upper limit cannot be directly transformed into an upper limit for λ_S without knowledge of λ_B since $\lambda_S = \lambda_B(\xi_S - 1)$.

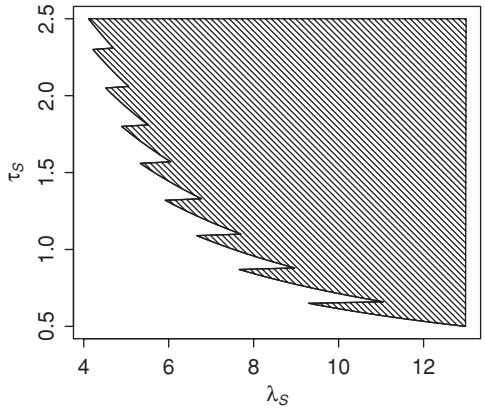

Figure 12. Dependence of the upper limit on the exposure time. The shaded area above and to the right of the curve is where the detection probability exceeds $\beta_{\min} = 0.90$. Thus, the curves give the upper limit (on the horizontal scale) as a function of the exposure time. The plot was made with $\lambda_B = 3$, $r = 5$, $\tau_S = \tau_B$, and $\alpha = 0.05$ and shows how the upper limit generally decreases as the exposure time increases. Because of the discrete nature of Poisson data, the probability of type one error can not be set exactly equal to α. This results in the step function nature of α in Figure 2 and the non-monotonic decrease of the upper limit as a function of exposure time here.

REFERENCES

Avni, Y., Soltan, A., Tananbaum, H., & Zamorani, G. 1980, ApJ, 238, 800
Ayres, T. 1999, ApJ, 525, 1009
Ayres, T. R. 2004, ApJ, 608, 957
Calderwood, T., Dobrzycki, A., Jessop, H., & Harris, D. E. 2001, in ASP Conf. Ser. 238, Astronomical Data Analysis Software and Systems X, ed. F. R. Harnden, F. A. Primini, & H. E. Payne (San Francisco, CA: ASP), 443
Carrera, F. J., et al. 2007, A&A, 469, 27
Casella, G., & Berger, R. L. 2002, Statistical Inference (2nd ed.; Pacific Grove, CA: Duxbury Press)
Damiani, F., Maggio, A., Micela, G., & Sciortino, S. 1997, ApJ, 483, 350
Dobrzycki, A., Jessop, H., Calderwood, T. J., & Harris, D. E. 2000, BAAS, 32, 1228
Eadie, W. T., Drijard, D., James, F. E., Roos, M., & Sadoulet, B. 1971, Statistical Methods in Experimental Physics (Amsterdam: North-Holland)
Eddington, A. S. 1913, MNRAS, 73, 359
Erdeve, I., Kalemci, E., & Alpar, M. A. 2009, ApJ, 696, 1792
Feldman, G. J., & Cousins, R. D. 1998, Phys. Rev. D, 57, 3873

Freeman, P. E., Kashyap, V., Rosner, R., & Lamb, D. Q. 2002, ApJS, 138, 185
Garwood, F. 1936, Biometrika, 28, 437
Gehrels, N. 1986, ApJ, 303, 335
Gelman, A., Meng, X.-L., & Stern, H. 1996, Statistica Sinica, 6, 733
Gilman, D. A., Hurley, K. C., Seward, F. D., Schnopper, H. W., Sullivan, J. D., & Metzger, A. E. 1986, ApJ, 300, 453
Gradshteyn, I. S., & Ryzhik, I. M. 1980, Table of Integrals, Series, and Products (4th ed.; San diego, CA: Academic Press)
Harnden, F. R., Jr., Fabricant, D., Harris, D., & Schwartz, D. 1984, SAO Spc. Rep. 393
Hughes, J. P., Chugai, N., Chevalier, R., Lundquist, P., & Schlegel, E. 2007, ApJ, 670, 1260
Isobe, T., Feigelson, E. D., & Nelson, P. I. 1986, ApJ, 306, 490
Kashyap, V., et al. 2008, HEAD, 10.0302
Kraft, R. P., Burrows, D. N., & Nousek, J. A. 1991, ApJ, 374, 344
Loewenstein, M., Kusenko, A., & Biermann, P. L. 2009, ApJ, 700, 426
Mandelkern, M. 2002, Stat.Sci., 17, 149
Loredo, T. 1992, in Statistical Challenges in Modern Astronomy, ed. E. D. Feigelson & G. J. Babu (New York: Springer-Verlag), 275
Marshall, H. 1992, in Statistical Challenges in Modern Astronomy I, ed. E. D. Feigelson & G. J. Babu (Berlin: Springer), 247
Mattox, J. R., et al. 1996, ApJ, 461, 396
Park, T., van Dyk, D. A., & Siemiginowska, A. 2008, ApJ, 688, 807
Pease, D. O., Kashyap, V. L., & Drake, J. J. 2006, ApJ, 636, 436
Perez-Torres, M. A., Zandanel, F., Guerrero, M. A., Pal, S., Profumo, S., Prada, F., & Panessa, F. 2009, MNRAS, 396, 2237
Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V., & Siemiginowska, A. 2002, ApJ, 571, 545
Rudnick, L., & Lemmerman, J. A. 2009, ApJ, 697, 1341
Schlegel, E. M., & Petre, R. 1993, ApJ, 412, L29
van Dyk, D. A., Connors, A., Kashyap, V. L., & Siemiginowska, A. 2001, ApJ, 548, 224
Vikhlinin, A., Forman, W., & Jones, C. 1997, ApJ, 474, L7
Weisskopf, M. C., Wu, K., Trimble, V., O'Dell, S. L., Elsner, R. F., Zavlin, V. E., & Kouvelioutou, C. 2007, ApJ, 657, 1026