Astronomical Data Analysis Software and Systems VI
ASP Conference Series, Vol. 125, 1997
Gareth Hunt and H. E. Payne, eds.

Heuristic Estimates of Weighted Binomial Statistics for Use in Detecting Rare Point Source Transients

James Theiler and Jeff Bloch

Astrophysics and Radiation Measurements Group, MS-D436 Los Alamos National Laboraotry, Los Alamos, NM 87545 e-mail: jt@lanl.gov, jbloch@lanl.gov

Abstract. The ALEXIS[1] (Array of Low Energy X-ray Imaging Sensors) (Priedhorsky et al. 1989) satellite scans nearly half the sky every fifty seconds, and downlinks time-tagged photon data twice a day. The standard science quicklook processing produces over a dozen sky maps at each downlink, and these maps are automatically searched for potential transient point sources. We are interested only in *highly significant* point source detections, and, based on earlier Monte-Carlo studies (Roussel-Dupré et al. 1996), only consider $p < 10^{-7}$, which is about 5.2 "sigmas." Our algorithms are therefore required to operate on the far tail of the distribution, where many traditional approximations break down. Although an exact solution is available for the case of unweighted counts (Lampton 1994), the problem is more difficult in the case of weighted counts. We have found that a heuristic modification of a formula derived by Li & Ma (1983) provides reasonably accurate estimates of p-values for point source detections even for very low p-value detections.

1. Introduction

We test the null hypothesis of no point source (assuming a spatially uniform background) at a given location by enclosing that location with a source kernel (whose area $A_{\rm src}$ is generally matched to the point-spread-function of the telescope) and then enclosing the source kernel with a relatively large background annulus (area $A_{\rm bak}$). Given $N_{\rm src}$ photons in the source kernel, and $N_{\rm bak}$ photons in the background annulus, the problem is to determine whether the number of source photons is *significantly* larger than expected under the null.

More sensitive point source detection is obtained by weighting the photons to match the point-spread function of the telescope more precisely. Further enhancements are obtained for ALEXIS data by weighting also according to instantaneous scalar background rate, pulse height, and position on the detector. In this case, we ask whether the weighted sum of photons in the source region is significantly larger than expected under the null.

[1] http://nis-www.lanl.gov/nis-projects/alexis/

2. Unweighted Counts

If counts are unweighted (i.e., all weights are equal), then it is possible to write down an exact, explicit expression for the probability of seeing $N_{\rm src}$ or more photons in the source kernel, assuming $N_{\rm total} = N_{\rm src} + N_{\rm bak}$ is fixed. This is a binomial distribution, and Lampton (1994) showed that the p-value associated with this observation can be expressed in terms of the incomplete beta function: $p = I_f(N_{\rm src}, N_{\rm bak}+1)$, where $f = A_{\rm src}/(A_{\rm src}+A_{\rm bak})$. See also Alexandreas et al. (1994), for an alternative derivation of an equivalent expression (the assumption that $N_{\rm total}$ is fixed is replaced by a Bayesian argument).

If the count rate is high (or the exposure long), so that $N_{\rm src}$ and $N_{\rm bak}$ are large, then an appropriate Gaussian approximation can be used. In general, this involves finding a "signal" and dividing it by the square root of its variance.

Case 1u. The most straightforward approach uses the signal $N_{\rm src} - \alpha N_{\rm bak}$, where $\alpha = A_{\rm src}/A_{\rm bak}$. Under the null hypothesis, this signal has an expected value of zero, and a variance—if $N_{\rm src}$ and $N_{\rm bak}$ are treated as independent Poisson sources—of $N_{\rm src} + \alpha^2 N_{\rm bak}$. To get a p-value, use

$$p = \mathcal{S}\left(\frac{N_{\rm src} - \alpha N_{\rm bak}}{\sqrt{N_{\rm src} + \alpha^2 N_{\rm bak}}}\right), \tag{1}$$

where $\mathcal{S}(s) = \frac{1}{2}(1 - \mathrm{erfc}(s/\sqrt{2}))$ converts "sigmas" of significance into a one-tailed p-value.

Case 2u. An alternative approach, suggested by Li & Ma (1983), treats the sum $N_{\rm total} = N_{\rm src} + N_{\rm bak}$, as fixed, so that $N_{\rm src}$ and $N_{\rm bak}$ are binomially distributed. In particular, choose the signal $N_{\rm src} - fN_{\rm total}$, and note that the variance of $N_{\rm src}$ is given by $f(1-f)N_{\rm total}$, while the variance of $N_{\rm total}$ is by definition zero. In that case

$$p = \mathcal{S}\left(\frac{N_{\rm src} - fN_{\rm total}}{\sqrt{f(1-f)N_{\rm total}}}\right) = \mathcal{S}\left(\frac{N_{\rm src} - \alpha N_{\rm bak}}{\sqrt{\alpha N_{\rm src} + \alpha N_{\rm bak}}}\right). \tag{2}$$

Case 3u. By looking at a ratio of Poisson likelihoods, Li & Ma (1983) also derived a more complicated equation

$$p \;=\; \mathcal{S}\left(\sqrt{2\left\{N_{\rm src}\ln(N_{\rm src}/\hat{N}_{\rm src}) + N_{\rm bak}\ln(N_{\rm bak}/\hat{N}_{\rm bak})\right\}}\right), \tag{3}$$

where $\hat{N}_{\rm src} = fN_{\rm total}$ and $\hat{N}_{\rm bak} = (1-f)N_{\rm total}$. This is considerably more accurate than Eqs. (1,2) when $N_{\rm src}$ and $N_{\rm bak}$ are not large, but is still just an approximation to Lampton's exact formula. Abramowitz & Stegun (1972) provide several approximations to the incomplete beta function, one of which (25.5.19) is an asymptotic series whose first term looks very much like the Li & Ma formula. The left panel of Figure 1 compares these cases, along with the Lampton (1994) formula, using a Monte-Carlo simulation.

3. Weighted Counts

Define $W_{\rm src} = \sum_{i\in{\rm src}} w_i$ and $Q_{\rm src} = \sum_{i\in{\rm src}} w_i^2$, where w_i is the weight of the i-th photon. Notice that when all the weights are equal to one, we have $Q_{\rm src} =$

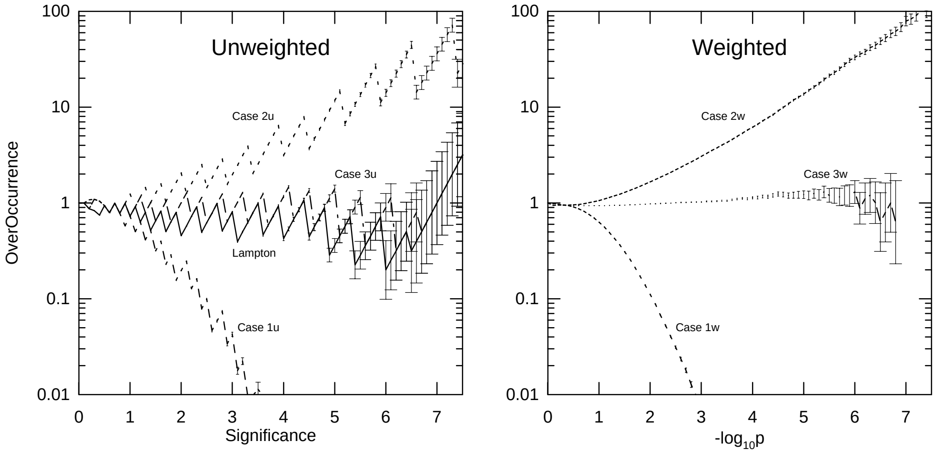

Figure 1. Results of Monte-Carlo experiments with $N = 100$ photons, with $f = 0.1$, and with $T = 10^7$ trials. For the weighted experiment, N weights were uniformly chosen from zero to one, and assigned to the N photons. The photons were randomly assigned to the source kernel or background annulus with probabilities f and $1 - f$ respectively. Values of $W_{\rm src}$, $W_{\rm bak}$, $Q_{\rm src}$, and $Q_{\rm bak}$ were computed, and a p-value was computed using the formulas for the three cases. As the p-values were computed, a cumulative histogram $H(p)$ was built indicating the number of times a p-value less than p was observed. Since we expect $H(p) = pT$, we plotted $H(p)/pT$ as the frequency of "overoccurrence" of that p-value. The plot is this overoccurrence as a function of "significance," defined by $-\log_{10} p$.

$W_{\rm src} = N_{\rm src}$ and $Q_{\rm bak} = W_{\rm bak} = N_{\rm bak}$. Note also that $W_{\rm src}/N_{\rm src} = \langle w_i \rangle_{i \in {\rm src}}$, and that $Q_{\rm src}/W_{\rm src} = \langle w_i^2 \rangle / \langle w_i \rangle$. We do not make any assumptons about weights averaging or summing to unity. (We define $W_{\rm bak}$ and $Q_{\rm bak}$ similarly.)

Generalizing **Case 1u**, we define the signal as $W_{\rm src} - \alpha W_{\rm bak}$ and then treating source and background as independent, we can write the variance as $Q_{\rm src} + \alpha^2 Q_{\rm bak}$. We can similarly generalize **Case 2u** and obtain:

$$\textbf{Case 1w:} \qquad p = \mathcal{S}\left(\frac{W_{\rm src} - \alpha W_{\rm bak}}{\sqrt{Q_{\rm src} + \alpha^2 Q_{\rm bak}}} \right). \tag{4}$$

$$\textbf{Case 2w:} \qquad p = \mathcal{S}\left(\frac{W_{\rm src} - \alpha W_{\rm bak}}{\sqrt{\alpha Q_{\rm src} + \alpha Q_{\rm bak}}} \right). \tag{5}$$

Case 3w: It is not as straightforward to generalize Eq. (3), but we have tried the following heuristic:

$$p = \mathcal{S}\left(\sqrt{ \left(\frac{2W_{\rm total}}{Q_{\rm total}} \right) \left\{ W_{\rm src} \ln(W_{\rm src}/\hat{W}_{\rm src}) + W_{\rm bak} \ln(W_{\rm bak}/\hat{W}_{\rm bak}) \right\} } \right), \tag{6}$$

where $\hat{W}_{\rm src} = f W_{\rm total}$ and $\hat{W}_{\rm bak} = (1 - f) W_{\rm total}$. The Monte-Carlo results shown in Figure 1 indicate that this heuristic provides reasonably accurate p-values even for very small values of p.

4. Limit of Precisely Known Background

An interesting limit occurs as the background annulus becomes large. Here, $A_{\mathrm{bak}} \to \infty$, and the expected backgrounds $\hat{N}_{\mathrm{src}}$, $\hat{W}_{\mathrm{src}}$, etc. are all precisely known. For the unweighted counts, the exact p-value can be expressed in terms of the incomplete gamma function: $p = 1 - \Gamma(N_{\mathrm{src}}, \hat{N}_{\mathrm{src}})/\Gamma(N_{\mathrm{src}})$. The Gaussian estimate of significance is straightforward[2] both for the unweighted case, $p = \mathcal{S}\left(\frac{N_{\mathrm{src}} - \hat{N}_{\mathrm{src}}}{\sqrt{\hat{N}_{\mathrm{src}}}}\right)$, and for the weighted case: $p = \mathcal{S}\left(\frac{W_{\mathrm{src}} - \hat{W}_{\mathrm{src}}}{\sqrt{\hat{Q}_{\mathrm{src}}}}\right)$. In this limit, Eq. (6) becomes

$$p = \mathcal{S}\left(\sqrt{2\left(\hat{W}_{\mathrm{src}}/\hat{Q}_{\mathrm{src}}\right)\left(W_{\mathrm{src}} \ln(W_{\mathrm{src}}/\hat{W}_{\mathrm{src}}) - (W_{\mathrm{src}} - \hat{W}_{\mathrm{src}})\right)}\right). \qquad (7)$$

Marshall (1994) has suggested an empirical formula $p = \mathcal{S}\left(\frac{W_{\mathrm{src}} - \hat{W}_{\mathrm{src}} + \Delta}{\sqrt{\hat{Q}_{\mathrm{src}} + \Delta}}\right)$, where $\Delta = 0.7\hat{Q}_{\mathrm{src}}/\hat{W}_{\mathrm{src}}$, which produced reasonable results in his simulations, but does not appear well suited for p-values at the far tail of the distribution.

Acknowledgments. This work was supported by the United States Department of Energy.

References

Abramowitz, M., & Stegun, I. A. 1972, Handbook of Mathematical Functions (Dover, New York), 945

Alexandreas, D. E., et al. 1993, Nucl. Instr. Meth. Phys. Res. A328, 570

Babu, G. J., & Feigelson, E. D. 1996, Astrostatistics (Chapman & Hall, London), 113

Lampton, M. 1994, ApJ, 436, 784

Li, T.-P., & Ma, Y.-Q. 1983, ApJ, 272, 317

Marshall, H. L. 1994, in ASP Conf. Ser., Vol. 61, Astronomical Data Analysis Software and Systems III, ed. D. R. Crabtree, R. J. Hanisch & J. Barnes (San Francisco: ASP), 403

Priedhorsky, W. C., Bloch, J. J., Cordova, F., Smith, B. W., Ulibarri, M., Chavez, J., Evans, E., Seigmund, O., H. W., Marshall, H., & Vallerga, J. 1989, in Berkeley Colloquium on Extreme Ultraviolet Astronomy, Berkeley, CA, vol 2873, 464

Roussel-Dupré, D., Bloch, J. J., Theiler, J., Pfafman, T., & Beauchesne, B. 1996, in ASP Conf. Ser., Vol. 101, Astronomical Data Analysis Software and Systems V, ed. G. H. Jacoby & J. Barnes (San Francisco: ASP), 112

---

[2]Babu & Feigelson (1996) incorrectly suggest $p = \mathcal{S}\left((N_{\mathrm{src}} - \hat{N}_{\mathrm{src}})/\sqrt{N_{\mathrm{src}} + \hat{N}_{\mathrm{src}}}\right)$.