DRAFT 1.1

30 May, 2012

Discovery sensitivity for a counting experiment with background uncertainty

Glen Cowan

Physics Department, Royal Holloway, University of London, Egham, TW20 0EX, U.K.

Abstract

An approximate expression is derived for the expected discovery significance for a Poisson counting experiment in which the background rate is constrained by a Poisson control measurement. The formula is based on a test statistic using the profile likelihood ratio and the expected significance is approximated by using the Asimov data set, as outlined in Ref. [1]. The validity of the new expression is compared with Monte Carlo results and to the other formulae for expected significance often used in particle physics.

1 Introduction

In a search for a new particle physics process one may observe a certain number number of events, n, modeled as following a Poisson distribution with a mean of $s + b$, where s and b represent the expected numbers of events from the (new) signal process and background processes, respectively. Here this will be referred to as a Poisson counting experiment. To establish a discovery of the signal, one can calculate the p-value of the hypothesis that $s = 0$, or equivalently it is convenient to use the Gaussian significance, Z. This is related to the p-value by

$$Z = \Phi^{-1}(1 - p) \, , \tag{1}$$

where Φ^{-1} is the quantile of the standard Gaussian (inverse of the cumulative distribution). For the case where the p-value refers to the background-only ($s = 0$) hypothesis, we will refer to the corresponding Z value as the discovery significance. In particle physics a widely used threshold for a discovery has been a p-value of 2.9×10^{-7} or less, corresponding to a significance of $Z = 5$ or more.

When designing a new experiment it is important to know what discovery significance to expect if a certain signal model is in fact true. For this one can report the mean or median value of Z under assumption of some nominal value of s, i.e., assuming that n will have a mean of $s + b$. As noted in Ref. [1], because the p-value and significance Z have a nonlinear, monotonic relation, it is convenient to take "expected significance" to refer to the median, so that the median Z is given by Z evaluated with the median p.

In Sec. 2 the basic formalism of quantifying discovery significance with a statistical test is reviewed. An often used measure of expected discovery significance for the Poisson counting experiment is given by $s/\sqrt{b}$. In Sec. 3 we examine the rationale behind this formula and discuss its extension to the case where the value of b is uncertain. In Ref. [1], an improvement over $s/\sqrt{b}$ was derived for the case of Poisson distributed n, and it was shown that the new expression is a better approximation to the true median discovery significance especially when the condition $s \ll b$ does not hold. For completeness this is reviewed in Sec. 4. The primary result of this note is an extension of the Poisson-based formula for expected discovery significance to the case where b is uncertain, and is given in Eq. (20) of Sec. 5. Conclusions are given in Sec. 6.

2 Discovery as a statistical test

In particle physics one frequently quantifies the significance of an observed signal by quoting the p-value of the background-only hypothesis, i.e., that of $s = 0$. One method for defining the p-value for a hypothesized value of s has been to construct a test statistic based on the profile likelihood ratio,

$$\lambda(s) = \frac{L(s, \hat{\hat{\boldsymbol{\theta}}}(s))}{L(\hat{s}, \hat{\boldsymbol{\theta}})} \, . \tag{2}$$

Here $L(s, \boldsymbol{\theta})$ is the likelihood function that represents the probability for the measurement (i.e., the number of events n plus any subsidiary measurements), under assumption of the signal parameter s and any additional (nuisance) parameters $\boldsymbol{\theta}$. The double-hat notation in the numerator of Eq. (2) refers to the values of $\boldsymbol{\theta}$ that maximize the likelihood under

assumption of the specified value of s, and the single hats in the denominator refer to the values that give the unconditional maximum of the likelihood (the ML estimators). The numerator is thus the profile likelihood; the denominator is the maximum of the likelihood.

In Ref. [1] the profile likelihood ratio was used as the basis of a test statistic defined as

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{s} \geq 0\ , \\ 0 & \hat{s} < 0\ . \end{cases} \tag{3}$$

This is defined so that large values of q_0 correspond to increasing disagreement between the data and the hypothesized value of $s = 0$. Although we consider here only the case where a physical signal model has $s > 0$, the estimator $\hat{s}$ is defined as the value that maximizes the likelihood even if this is negative. This occurs if the number of observed events is smaller than the expected background. The statistic q_0 is defined as zero for $\hat{s} < 0$ so that it reflects a discrepancy between the data and hypothesis only in the case where the observed signal rate is positive.

3 $s/\sqrt{b}$ and related measures of discovery sensitivity

In particle physics the quantity $s/\sqrt{b}$ has been widely used as a measure of expected discovery significance. The rationale behind this formula is that a Poisson distributed quantity n with a large mean value $s+b$ can be approximated by a Gaussian distributed variable x with mean $s + b$ and standard deviation $\sqrt{s + b}$. The p-value of the background-only hypothesis given an observation x is therefore

$$p = 1 - \Phi\left(\frac{x-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{x-b}{\sqrt{b}}\right)\ , \tag{4}$$

where $\mu = b$ and $\sigma = \sqrt{b}$ refer to the mean and standard deviation of x under assumption of $s = 0$. Using this with Eq. (1) gives the discovery significance

$$Z = \frac{x-b}{\sqrt{b}}\ . \tag{5}$$

The median (here equal to the mean) Z under assumption of a given value of s is therefore

$$\mathrm{med}[Z|s] = \frac{s}{\sqrt{b}}\ . \tag{6}$$

The intuitive explanation of this formula is that the standard deviation of n assuming background only is $\sqrt{b}$, and therefore the ratio $s/\sqrt{b}$ represents the size of the signal divided by the statistical error on n expected assuming signal is absent.

Often the expected number of background events is not known exactly but has some systematic uncertainty characterized by a standard deviation σ_b. In this case, one may generalize Eq. (6) to account both the statistical and systematic error in b by replacing $\sqrt{b}$ by the the quadratic sum of $\sqrt{b}$ and σ_b, so that the median significance becomes

$$\mathrm{med}[Z|s] = \frac{s}{\sqrt{b + \sigma_b^2}}\ . \tag{7}$$

This formula has been used in particle physics as a measure of expected discovery significance for processes where the uncertainty in the expected background cannot be neglected. In Sec. 5 a formal justification for Eq. (7) will be given and the limits of its validity investigated.

4 Poisson case with known background

If the expected number of background events, b, is known with negligible uncertainty, then the likelihood function for the Poisson counting experiment is

$$L(s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \; . \tag{8}$$

In Ref. [1] this problem was investigated using the test statistic q_0 as defined in Eq. (3). For a sufficiently large expected number of events, one can show using Wilks' theorem (see, e.g., [1] and references therein) that the discovery significance can be approximated by $Z = \sqrt{q_0}$. For the present problem this gives

$$Z = \sqrt{2\left(n \ln \frac{n}{b} + b - n\right)} \tag{9}$$

for $n > b$ and $Z = 0$ otherwise. It was also shown in Ref. [1] that one can approximate the median significance by replacing the data by the corresponding expectation values (the so-called Asimov data set). Substituting $s + b$ for n in Eq. (9) thus gives the Asimov approximation for the median significance, denoted here by $Z_{\rm A}$:

$$Z_{\rm A} = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)} \; . \tag{10}$$

Expanding the logarithm in s/b one finds

$$Z_{\rm A} = \frac{s}{\sqrt{b}} \left(1 + \mathcal{O}(s/b)\right) \; . \tag{11}$$

Thus the full expression for $Z_{\rm A}$ in Eq. (10) reduces to the widely used formula $s/\sqrt{b}$ in the limit $s \ll b$.

Although $s/\sqrt{b}$ has been widely used for expected discovery significance in cases where $s + b$ is large, one sees here that this approximation is strictly valid only for $s \ll b$.

Median values of the expected discovery significance Z for different values of s and b are shown in Fig. 1 (from [1, 2]). The solid curve shows Eq. (10), the dashed curve gives the approximation $s/\sqrt{b}$, and the points are the exact median values from Monte Carlo. The structure seen in the points is due to the discrete nature of the data. One sees that Eq. (10) provides a much better approximation to the true median than does $s/\sqrt{b}$ in regions where $s \ll b$ does not hold.

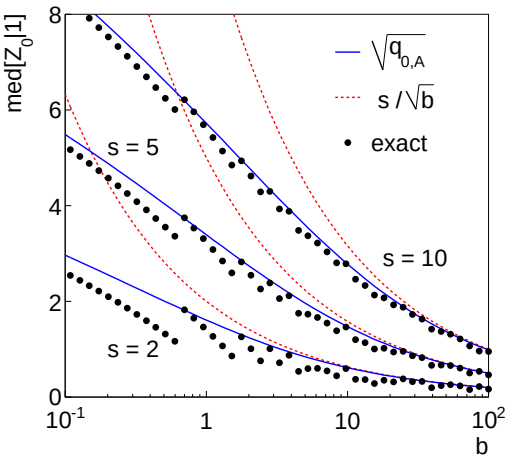

Figure 1: The median, assuming a mean number of signal events s, of the discovery significance Z for different values of s and b (from [1, 2]; see text).

5 Poisson case with uncertain background

If the expected number of background events, b, is not known one must treat it as a nuisance parameter in the likelihood function. Because b could be adjusted to accommodate any observed number of events, it would be impossible to reject the hypothesis of $s = 0$ unless some additional information is introduced that constrains b. Often this is done by means of a control measurement by counting the number of events m in a data sample where signal events are believed to be absent, and where the mean number of events can be related to the number of background events in the primary measurement of n. For example, one may take m as following a Poisson distribution with a mean of τb, where τ is a scale factor that we take here as known with negligible uncertainty.

The full likelihood function that describes both the primary measurement n and the control measurement m is therefore the product of the two corresponding Poisson distributions:

$$L(s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \, \frac{(\tau b)^m}{m!} e^{-\tau b} \; . \tag{12}$$

This problem has been studied in both particle physics and astrophysics (see, e.g., [1, 3, 4, 5]). To find the profile likelihood ratio one needs the estimators for b and s as well as the conditional estimator for b given a value of s:

$$\hat{s} = n - m/\tau \; , \tag{13}$$

$$\hat{b} = m/\tau \; , \tag{14}$$

$$\hat{\hat{b}}(s) = \frac{n + m - (1+\tau)s + \sqrt{(n + m - (1+\tau)s)^2 + 4(1+\tau)sm}}{2(1+\tau)} \; . \tag{15}$$

For the statistic q_0 one needs in particular $\hat{\hat{b}}(0)$, which from Eq. (15) is

$$\hat{\hat{b}}(0) = \frac{n+m}{1+\tau} \; . \tag{16}$$

As in Sec. 4 we use the approximation $Z = \sqrt{q_0}$, valid in the large sample limit, which gives

$$Z = \left[-2 \left(n \ln \left[\frac{n+m}{(1+\tau)n} \right] + m \ln \left[\frac{\tau(n+m)}{(1+\tau)m} \right] \right) \right]^{1/2} \tag{17}$$

for $n > \hat{b}$ and $Z = 0$ otherwise. This is the same as Eq. (17) of Ref. [3] and Eq. (25) of Ref. [5].

As in Sec. 4 we replace the data values n and m by their expectation values $s + b$ and τb to give the "Asimov" approximation for the median significance. Making this substitution in Eq. (17) gives

$$Z_{\mathrm{A}} = \left[-2 \left((s+b) \ln \left[\frac{s + (1+\tau)b}{(1+\tau)(s+b)} \right] + \tau b \ln \left[1 + \frac{s}{(1+\tau)b} \right] \right) \right]^{1/2} . \tag{18}$$

The case where the control measurement m has a small relative statistical uncertainty corresponds to τ very large, and in this limit Eq. (18) reverts to the expression for known b given by Eq. (10).

It is useful to re-express Eq. (18) in terms of the uncertainty one would quote on the background on the basis of the control measurement m. The estimator for b is given by Eq. (14), and because the variance of m is equal to its mean, τb, the variance of $\hat{b}$ is

$$V[\hat{b}] \equiv \sigma_b^2 = \frac{b}{\tau} . \tag{19}$$

Using this equation to eliminate τ from (18) gives the result

$$Z_{\mathrm{A}} = \left[2 \left((s+b) \ln \left[\frac{(s+b)(b+\sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b+\sigma_b^2)} \right] \right) \right]^{1/2} . \tag{20}$$

By expanding this expression in powers of s/b and σ_b^2/b one finds

$$Z_{\mathrm{A}} = \frac{s}{\sqrt{b + \sigma_b^2}} \left(1 + \mathcal{O}(s/b) + \mathcal{O}(\sigma_b^2/b) \right) . \tag{21}$$

One sees that that the expression given in Eq. (7) and justified on intuitive grounds results from the significance based on Wilks' theorem and use of the Asimov data set. The simple formula given by Eq. (21) is expected to be valid in cases where one has $s \ll b$ and $\sigma_b^2 \ll b$. From Eq. (19) we have $\sigma_b^2/b = 1/\tau$, and because the expectation value of m is $E[m] = \tau b$, the requirement $\sigma_b^2 \ll b$ is equivalent to $E[m] \gg b$ or $\tau \gg 1$. That is, the expected number of events in the control region should be large compared to the expected number of background events contributing to the primary measurement of n (and in addition $s \ll b$ must hold).

Figure 2 shows the median discovery significance for $s = 2, 5$ and 10 as a function of b for (left) several values of σ_b/b and (right) several values of τ. In each plot the upper set of curves (points) corresponds to the smaller σ_b/b or larger τ. The points are based on Eq. (17) and computed with Monte Carlo. The structure in the points is due to the discreteness of the Poisson distributed data. The dashed and solid curves show the predictions of Eqs. (7) and (20), respectively. Although both of these formulae agree with the Monte Carlo values for sufficiently large b, the full formula from Eq. (20) is clearly in far better agreement for low b. This is understandable because for decreasing b the ratios s/b and σ_b^2/b become large and thus the approximation of Eq. (7) is expected to break down.

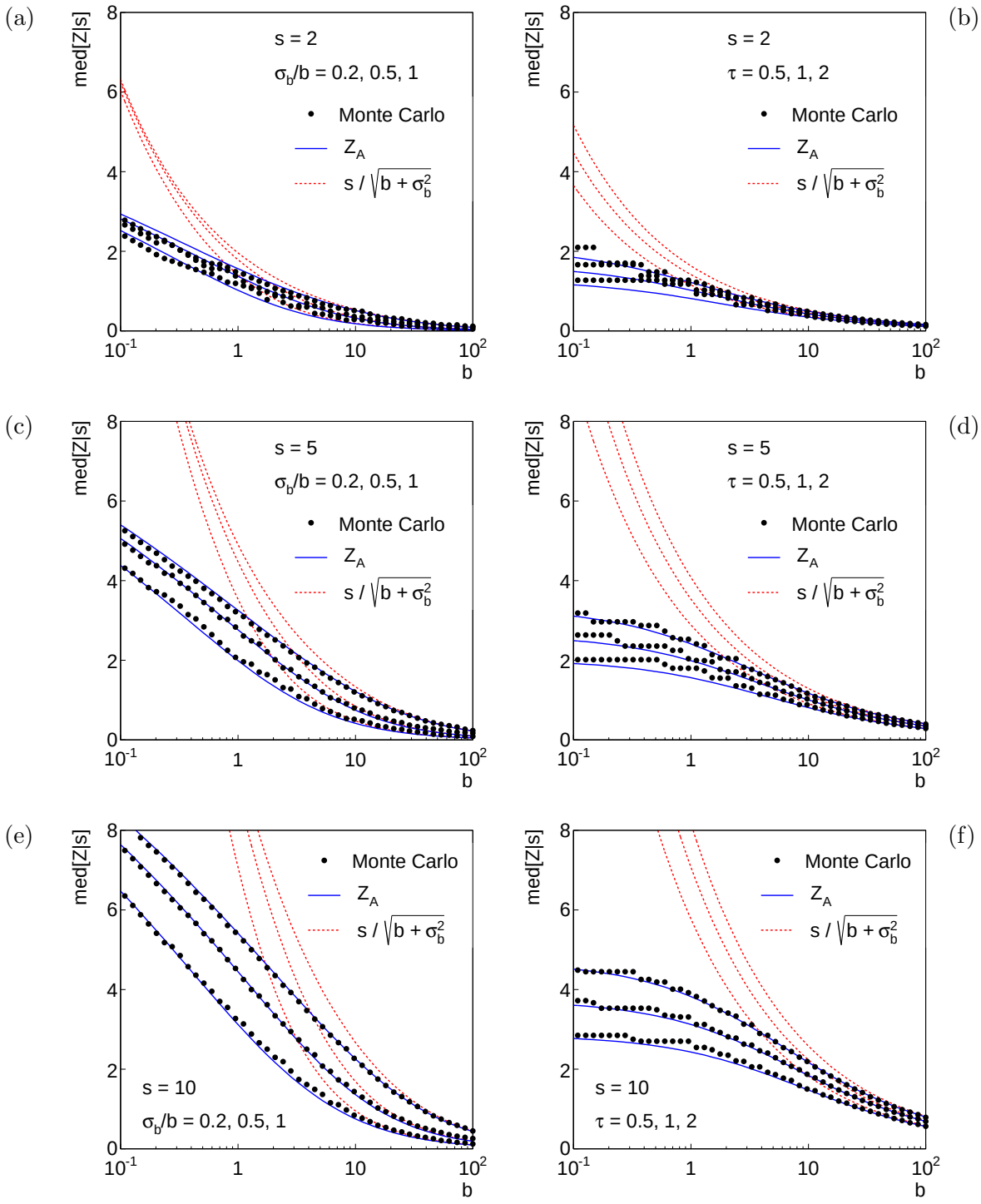

Figure 2: The median, assuming an expected number of signal events $s = 2, 5$ and 10 of the discovery significance Z as a function of the expected number of background events b for (left) several values of σ_b/b and (right) several values of τ (see text).

6 Conclusions

A simple expression for the expected discovery significance is given by Eq. (20), which can be used for a counting experiment in which the expected number of background events is uncertain. Specifically the formula treats the case where the expected background is constrained by a control measurement consisting of a Poisson distributed value. In the limit of small background uncertainty and also a small signal rate ($s \ll b$), Eq. (20) reduces to Eq. (7), which has been used in particle physics. The numerical studies shown, however, indicate that for important ranges of background rate and uncertainty, the limiting formula severely overestimates the expected discovery significance, whereas the full formula is in very good agreement with exact Monte Carlo results.

References

[1] Glen Cowan, Kyle Cranmer, Eilam Gross and Ofer Vitells, Eur. Phys. J. C 71 (2011) 1554.

[2] Glen Cowan, *Use of the profile likelihood function in searches for new physics* in H. Prosper and L. Lyons (eds.), Proceedings of the PHYSTAT 2011 Workshop, CERN-2011-006 (2011), p. 109. Figure 1(a) in this paper corrects a minor numerical error in Fig. 7 of Ref. [1].

[3] Tipei Li and Yuqian Ma, Astrophysical Journal 272 (1983) 317–324.

[4] Kyle Cranmer, *Statistical Challenges for Searches for New Physics at the LHC*, proceedings of PhyStat2005, Oxford; arXiv:physics/0511028.

[5] Robert D. Cousins, James T. Linnemann and Jordan Tucker, NIM A 595 (2008) 480–501; arXiv:physics/0702156.