Glen Cowan, Eilam Gross
ATLAS Statistics Forum
8 May, 2008

Discovery significance with statistical uncertainty in the background estimate

1 Introduction

In a search for a new type of event, data samples are used to estimate the expected number events for signal s and background b for a given integrated luminosity. In the design phase of the search this is done with Monte Carlo (MC). Establishing a discovery requires rejecting the hypothesis of background only. To quantify the statistical significance of a discovery, one can quote the equivalent fluctuation Z of a standard Gaussian variable. In the limit $s \ll b$ and for sufficiently large b one finds $Z = s/\sqrt{b}$, where s and b are assumed to be known precisely.

In practice s and b are estimated using a finite amount of data, and therefore their values have statistical uncertainties. It can happen, for example, that no background events pass the cuts (e.g., in control regions or from background MC) so that the naive estimate for the background is $b = 0$. In such cases, expressions such as $s/\sqrt{b}$ clearly do not yield a sensible result. One might try, for example, to use an upper limit for b at some confidence level (CL). Which CL to choose is not obvious, however, and therefore at best this leads to an approximate recipe.

Here we describe methods for incorporating statistical uncertainties into an estimate of the discovery significance. For discovery, only the statistical uncertainty of the background estimate is relevant. For setting limits, the error on s enters as well. In this note we focus on discovery and treat s as known. To include the uncertainty on s is a straightforward generalization of the method described here.

The statistical formalism is based on the *profile likelihood ratio* as used in [1] and described in, e.g., [2]. For completeness the relevant ingredients are summarized here.

The profile likelihood ratio is a well-established tool for including systematic uncertainties that can be related to nuisance parameters (any parameter that is not of direct interest in the final result), such as background rates. It is also appropriate to use this method to include the statistical uncertainty resulting from a limited Monte Carlo sample into the expected significance of a future measurement. Once the actual measurement is carried out, the statistical uncertainty in the background comes no longer from Monte Carlo but rather from control regions using real data, but the mathematical treatment remains essentially the same.

In this note we consider a measurement consisting of a single number of events n found in a search region, combined with subsidiary measurements to determine the background. It is straightforward to extend this to measurements that include shapes of distributions or combinations of multiple channels.

2 Statistical formalism

Suppose n events are selected in a search region where both signal and background could be present. The expectation value of n can be written

$$E[n] = \mu s + b_{\rm tot} \ , \tag{1}$$

where s is the expected number from signal and $b_{\rm tot}$ is the expected total background (i.e., from all sources). Here μ is a strength parameter defined such that $\mu = 0$ corresponds to the background-only hypothesis and $\mu = 1$ gives the nominal signal rate plus background.

Suppose that $b_{\rm tot}$ consists of N components, i.e.,

$$b_{\rm tot} = \sum_{i=1}^{N} b_i \ . \tag{2}$$

To estimate the expected number of events from background component i using Monte Carlo, we generate a sample of M_i events, and in addition the generator calculates a cross section σ_i. From these we have the equivalent integrated luminosity of the MC sample, $L_i = M_i/\sigma_i$.

Suppose m_i of these events are selected in the search region. From a statistical standpoint, this is equivalent to having a subsidiary measurement m_i modeled as following a Poisson distribution with expectation value

$$E[m_i] = \tau_i b_i \ . \tag{3}$$

Here τ_i is a scale factor that relates the mean number of events that contribute to n (the primary measurement), to that of the ith subsidiary measurement. If m_i is the number of MC events found in the search region, then τ_i is the ratio of the integrated luminosity of the Monte Carlo sample to that of the data,

$$\tau_i = \frac{L_{{\rm MC},i}}{L_{\rm data}} \ . \tag{4}$$

In the case where the m_i represents a number of events found in a control region based on real data, the τ_i is effectively the ratio of the sizes of the control to signal regions. In either case we will assume that the τ_i can be determined with negligible uncertainty.

The likelihood function for the parameters μ and $\mathbf{b} = (b_1, \ldots, b_N)$ is the product of Poisson probabilities:

$$L(\mu, \mathbf{b}) = \frac{(\mu s + b_{\rm tot})^n}{n!} e^{-(\mu s + b_{\rm tot})} \prod_{i=1}^{N} \frac{(\tau_i b_i)^{m_i}}{m_i!} e^{-\tau_i b_i} \ . \tag{5}$$

Here μ is the parameter of interest; the components of $\mathbf{b}$ are nuisance parameters.

To test a hypothesized value of μ, one computes the profile likelihood ratio

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\mathbf{b}}})}{L(\hat{\mu}, \hat{\mathbf{b}})} \tag{6}$$

where the double-hat notation refers to the conditional maximum-likelihood estimators (MLEs) for the given value of μ, and the single hats denote the unconditional MLEs. In addition it is convenient to define

$$q_\mu = -2 \ln \lambda(\mu) \ . \tag{7}$$

The ratio $\lambda(\mu)$ is expected to be close to unity (i.e., q_μ is near zero) if the data are in good agreement with the hypothesized value of μ.

Suppose the data results in a value of $q_\mu = q_{\rm obs}$. The level of agreement between the data and hypothesized μ is given by the p-value,

$$p = \int_{q_{\rm obs}}^{\infty} f(q_\mu|\mu)\, dq_\mu \ , \tag{8}$$

where $f(q_\mu|\mu)$ is the sampling distribution of q_μ under the assumption of μ.

One can define the significance corresponding to a given p-value as the number of standard deviations Z at which a Gaussian random variable of zero mean would give a one-sided tail area equal to p. That is, the significance Z is related to the p-value by

$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z) \ , \tag{9}$$

where Φ is the cumulative distribution for the standard (zero mean, unit variance) Gaussian. Equivalently one has

$$Z = \Phi^{-1}(1 - p) \ , \tag{10}$$

where Φ^{-1} is the quantile of the standard Gaussian (inverse of the cumulative distribution).

In order to establish a discovery we try to reject the background-only hypothesis $\mu = 0$. For example, a threshold significance of $Z = 5$ corresponds to a p-value of 2.87×10^{-7}.

To find the p-value we need the sampling distribution $f(q_\mu|\mu)$ (specifically, for discovery we need $f(q_0|0)$). Under a set of regularity conditions and for a sufficiently large data sample, *Wilks' theorem* says that for a hypothesized value of μ, the pdf of the statistic $q_\mu = -2 \ln \lambda(\mu)$ approaches the chi-square pdf for one degree of freedom [3]. More generally, if there are N parameters of interest, i.e., those parameters that do not get a double hat in the numerator of the likelihood ratio (6), then q_μ asymptotically follows a chi-square distribution for N degrees of freedom. A proof and details of the regularity conditions can be found in standard texts such as [4].

In the searches considered here, the data samples are usually large enough to ensure the validity of the asymptotic formulae for the likelihood-ratio distributions. Nevertheless the distributions are modified because of constraints imposed on the expected number of events.

Usually when searching for a new type of particle reaction one regards the mean number of events contributed to any bin from any source, signal or background, to be greater than or equal to zero. In some analyses it could be a meaningful to consider a new effect that suppress the expected number of events, e.g., the presence of a new decay channel could mean that the number of decays to known channels is reduced. Here, however, we will regard any contribution to an expected number of events as non-negative.

Assuming only non-negative event rates, the maximum-likelihood estimators for the parameters are constrained, e.g., $\hat{\mu} \geq 0$. If the observed number of events is below the level predicted by the background alone, then the maximum of the likelihood occurs for $\mu = 0$, i.e., negative μ is not allowed. The likelihood ratio is then (see (6)),

$$\lambda(0) = \frac{L(0, \hat{\hat{\mathbf{b}}})}{L(\hat{\mu}, \hat{\mathbf{b}})} = \frac{L(0, \hat{\hat{\mathbf{b}}})}{L(0, \hat{\mathbf{b}})} = 1 \; , \tag{11}$$

since $\hat{\mu} = 0$ and therefore $\hat{\hat{\mathbf{b}}} = \hat{\mathbf{b}}$. The statistic $q_0 = -2 \ln \lambda(0)$ is therefore equal to zero.

Under the background-only hypothesis, the data will fall above or below the background expectation with approximately equal probability. In those cases where the data fluctuate up we have $\hat{\mu} > 0$ and q_0 follows a chi-square pdf for one degree of freedom, $f_{\chi^2_1}$. If $\hat{\mu} = 0$, then $q_0 = 0$. Assuming a fraction w for the cases with $\hat{\mu} > 0$ one has the pdf

$$f(q_0|0) = w f_{\chi^2_1}(q_0) + (1 - w)\delta(q_0) \; . \tag{12}$$

In the usual case where upward and downward fluctuations are equally likely we have $w = 1/2$.

Consider now the variable

$$u = \sqrt{q_0} = \sqrt{-2 \ln \lambda(0)} \; , \tag{13}$$

which has the pdf

$$f(u) = \Theta(u) w \sqrt{\frac{2}{\pi}} e^{-u^2/2} + (1 - w)\delta(u) \; , \tag{14}$$

where $\Theta(u) = 1$ for $u \geq 0$ and is zero otherwise. The second term in (14) follows from the fact that the values $q_0 = 0$ and $u = 0$ occur with equal probability, $1 - w$. Furthermore if a variable x follows the standard Gaussian, then one can show x^2 follows a chi-square distribution for one degree of freedom. Therefore if x^2 follows a χ^2 distribution, then $\sqrt{x^2}$ follows a Gaussian scaled up by a factor of two for $x > 0$ so as to have a total area of unity.

The p-value of the $\mu = 0$ hypothesis for a non-zero observation q_0 is therefore

$$p = P(u \geq \sqrt{q_0}) = 2w \int_{\sqrt{q_0}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \, du = 2w(1 - \Phi(\sqrt{q_0})) \; . \tag{15}$$

Combining this with equation (10) for the significance Z gives

$$Z = \Phi^{-1}(1 - 2w(1 - \Phi(\sqrt{q_0}))) \; . \tag{16}$$

In the usual case where the weights of the chi-square and delta-function terms are equal, i.e., $w = 1/2$, equation (16) reduces to to the simple formula

$$Z = \sqrt{q_0} = \sqrt{-2 \ln \lambda(0)} \; . \tag{17}$$

In cases where the data sample is not large enough to guarantee the validity of the asymptotic distribution, Monte Carlo studies can be carried out to find the sampling distribution of q_0, and from this the discovery significance for a given observed value.

To calculate the value of q_μ that one would obtain from a set of data values n and $m_1, \ldots, m_N$, one needs the unconditional estimators $\hat{\mu}$ and $\hat{\mathbf{b}}$, and the conditional MLEs $\hat{\hat{\mathbf{b}}}$, i.e., the values of $\mathbf{b}$ that maximize the likelihood for the specified value of μ.

The value of n itself is only available from the real data. To quantify the sensitivity of the analysis, one can report the expected or median significance one would obtain for data based on the best estimate of signal plus background. To good approximation, the median significance can be found by replacing n with $s + \hat{b}$. This is referred to as an 'Asimov' data set.[1]

In cases with more than one background component, it is easiest to solve for the required quantities numerically. A program for doing this is available from [6].

As an example consider a planned search [7] where six different background sources were investigated with separate MC samples. The expected numbers of events for a luminosity of $L = 1\,\mathrm{fb}^{-1}$ and the equivalent luminosity of the MC samples is shown Table 1.

Table 1: Number of expected background events b_i and equivalent luminosities L_i from Monte Carlo in a planned search.

b_i	L_i (fb^{-1})
11	0.95
0	2.67
1	2.98
0	1.22
0	2.98
0	0.75

For $L = 1\,\mathrm{fb}^{-1}$, $s = 312$ signal events are predicted. Using these numbers gives $Z = 18.1$. In a similar manner the 5σ discovery threshold is found to be at a luminosity of 72 pb^{-1}.

In this example, the impact of those background components where no events passed the cuts is small. If they are neglected entirely one finds $Z = 18.8$. If, however, the equivalent luminosity of one of the background samples had been much less than the data luminosity, then this would have a significant effect. Changing the luminosity of the last component in Table 1 from 0.75 to 0.075 results in $Z = 6.7$; if it is reduced to 0.0075, one finds $Z = 2.2$. A more detailed study of this effect is shown for the case of a single background component in Section 3.

3 Case of a single background component

If we only have one background component, i.e., a measurement m with mean τb, then the required estimators can be written easily in closed form. Taking into account the constraint $\hat{\mu} \geq 0$ one finds

[1] The name of the Asimov data set is inspired by the short story *Franchise*, by Isaac Asimov [5]. In it, elections are held by selecting a single voter to represent the entire electorate.

$$\hat{\mu} = \begin{cases} \frac{n-m/\tau}{s} & n \geq m/\tau \\ 0 & \text{otherwise} , \end{cases} \tag{18}$$

$$\hat{b} = \begin{cases} m/\tau & n \geq m/\tau \\ \frac{n+m}{\tau+1} & \text{otherwise} , \end{cases} \tag{19}$$

For the case of discovery, we are only interested in the hypothesis of $\mu = 0$. The conditional MLE for b given $\mu = 0$ is

$$\hat{\hat{b}} = \frac{n+m}{\tau+1} . \tag{20}$$

Putting together the ingredients for $\ln \lambda(0)$ yields

$$\ln \lambda(0) = \begin{cases} \psi(m, \tau\hat{\hat{b}}) + \psi(n, \hat{\hat{b}}) - \psi(m, \tau\hat{b}) - \psi(n, \hat{\mu}s + \hat{b}) & n \geq m/\tau , \\ 0 & \text{otherwise} , \end{cases} \tag{21}$$

where

$$\psi(x, y) = \begin{cases} 0 & x = y = 0 , \\ x \ln y - y & \text{otherwise} . \end{cases} \tag{22}$$

To find the median significance assuming the signal is present at the nominal rate, we replace n by $s + \hat{b}$ (the Asimov data set).

As an example where no background events survive the cuts, suppose $s = 7$, $\tau = 6.7$, and $m = 0$, and therefore we take $n = 7$ and have $\hat{b} = 0$. In this case the result simplifies to

$$q_0 = -2 \ln \lambda(0) = 2s \ln(1 + \tau) = 28.5 . \tag{23}$$

Using the asymptotic formula (17) for the significance gives

$$Z = \sqrt{q_0} = 5.3 . \tag{24}$$

The accuracy of this approximation can be checked over a range of values of b using a simple Monte Carlo simulation. Note that in this case because $m = 0$, the significance goes to zero as τ decreases to zero. That is, a very weak constraint on the background leads to a decreasing discovery significance.

In the limit where τ is very large, the background estimates $\hat{\hat{b}}$ and $\hat{b}$ both approach b, and equation (21) becomes

$$\ln \lambda(0) = -n \ln \frac{n}{b} + n - b . \tag{25}$$

for $n \geq b$ and zero otherwise. The median significance under the hypothesis of $\mu = 1$ can be approximated by substituting the 'Asimov' value $s + b$ for n in equation (25), i.e.,

$$\text{median}[Z|\mu = 1] \approx -(s+b)\ln\left(1+\frac{s}{b}\right) + s \tag{26}$$

The median discovery significance is for $\tau \to \infty$ (i.e., b known) therefore given by

$$Z = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)}\,. \tag{27}$$

If one then takes the limit $s \ll b$, then by expanding the logarithm in (26) and retaining terms up to order s^2, this becomes

$$\ln\lambda(0) \approx -\frac{s^2}{2b}\,. \tag{28}$$

Combining this with equation (17) for the significance then gives

$$Z \approx \frac{s}{\sqrt{b}}\,. \tag{29}$$

Thus in the limit of small s/b and with b well determined (large τ), one recovers the widely used formula.

Figure 1 shows the significance Z with $b = 10$ computed as a function of s. The plot shows the full calculation for Z for $\tau = 1$, the formula (27) valid for large τ, and the limiting formula (29) valid for large τ and $s \ll b$.

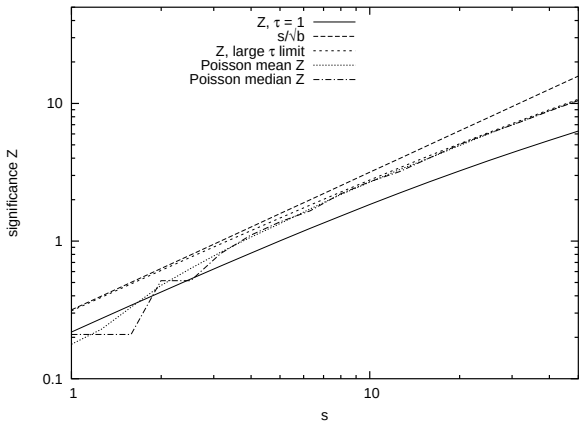

Figure 1: The significance Z as a function of the expected signal s according to several formulae (see text).

From the figure one can see that for $s = 10$, i.e., in this example $s/b = 1$, the approximate formula $s/\sqrt{b}$ gives 2.78, the full calculation gives 3.16, and the large τ approximation gives 1.84. For s/b much greater than unity, $s/\sqrt{b}$ overestimates the significance by an increasingly non-negligible amount. The effect of the statistical error on the estimate of b is also seen to be very significant through the substantial difference between the curves for $\tau = 1$ and the large-τ limit.

Also shown in Fig. 1 are curves for the mean and median significances computed numerically for the case of b known with n generated according to a Poisson distribution with mean $s+b$. These two curves represent the exact answer for the fixed b case in that they do not rely

on any asymptotic approximations. For significance values relevant to discovery or exclusion, say, $Z > 1$, they are in good agreement with the curve of equation (27) using the profile likelihood with Asimov data. For low s one can see that the profile likelihood prediction in the large τ limit is too high, but this is only in a region of very low significance values, not relevant for discovery or limits.

The significance calculation shown here can be used to help establish the appropriate amount of MC data needed to determine the discovery significance. Figure 2 shows the discovery significance Z as a function of the luminosity ratio $\tau = L_{\rm MC}/L_{\rm data}$ for several values of b and s.

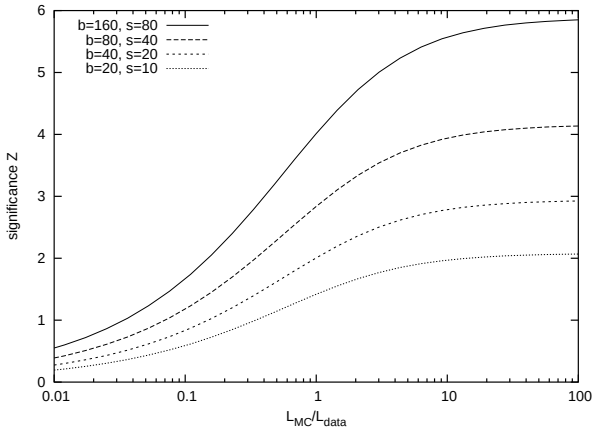

Figure 2: The significance Z as a function of the luminosity ratio $\tau = L_{\rm MC}/L_{\rm data}$ for several values of b and s.

In these examples one sees a rapid change in Z as the luminosity ratio τ varies between around 0.5 and 5. For $\tau < 0.5$ the significance is degraded by a factor of two; for $\tau > 5$ the improvement is slight.

References

[1] The ATLAS Statistics Forum, Statistical combination of ATLAS Higgs results, ATL-PHYS-PUB-2008-XXX, in preparation.

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

[3] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

[4] A. Stuart, J.K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A: *Classical Inference and the Linear Model* 6th Ed., Oxford Univ. Press (1999), and earlier editions by Kendall and Stuart.

[5] Isaac Asimov, *Franchise*, in *Isaac Asimov: The Complete Stories, Vol. 1*, Broadway Books, 1990.

[6] G. Cowan, `SigCalc`, a program for calculating discovery significance using profile likelihood, available from `www.pp.rhul.ac.uk/~cowan/stat/SigCalc/`.

[7] Christina Potter, private communication.