Glen Cowan, Eilam Gross
ATLAS Statistics Forum
2 December, 2009
DRAFT 1.0

Discussion and guidelines on significance

1 Introduction

In this note we discuss how to quantify the statistical significance of a discovery. We will address what measure of significance is appropriate in various circumstances, and discuss methods for including systematic uncertainties. The primary method which is sufficiently well established in HEP to the point where we feel justified in making concrete recommendations is based on the p-value of the background-only hypothesis, discussed below. Other related measures such as the Bayes factor can also be considered and contribute to the robustness of a claimed discovery.

In Section 2, the concept of a p-value is introduced along with the equivalent Gaussian significance (number of standard deviations). Section 3 applies these constructs to the case of a simple counting experiment, and validity of several approximations is explored. We also address the issue of the sensitivity of a planned experiment, which is effectively the expected (e.g., mean or median) significance with which one rejects the background-only hypothesis, under some assumption for the expected signal.

In Section 6 we address the issue of incorporating systematic uncertainties by using the profile likelihood ratio, and in addition the entire formalism is extended to the case of multiple channels, e.g., combination of multiple decay channels, or the use of multiple bins in a histogram.

2 p-values and equivalent Gaussian significance

The primary means by which one can establish a claim for New Physics is by rejecting the hypothesis that the observed data sample contains only Standard Model background events. The usual way of quantifying the level of compatibility (or lack thereof) between the data and a given hypothesis is to compute a p-value. This is the probability, under assumption of the hypothesis in question, of obtaining data with equal or lesser compatibility compared to the level found with the observed data.

Determining a p-value thus requires a definition of what data values constitue equal or lesser compatibility with the hypothesis in question, relative to the level actually observed. If we are testing the hypothesis that the events are all background, then this region is taken to be those data values that are equally signal-like or more so. In many analyses one can, at least approximately, choose the p-value to be constructed in such a way that one maximizes the probability of rejecting the background-only hypothesis if a particular signal model is true. For the examples treated below there will be a well-motivated choice for the definition of the p-value, and this choice is separate from the primary questions that we address in this note, which focuses on how best to approximate the significance in commonly occuring cases.

In addition to reporting the p-value, in HEP one often quotes the *significance*, defined 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 to the p-value. That is, the significance Z is related to the p-value by

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

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

$$Z = \Phi^{-1}(1-p)\ , \tag{2}$$

where Φ^{-1} is the quantile of the standard Gaussian (inverse of the cumulative distribution), which can be found using root with `TMath::NormQuantile`. The equivalent relation using the inverse error function is

$$Z = \sqrt{2}\mathrm{erf}^{-1}(1-2p)\ . \tag{3}$$

The relation between Z and p is illustrated in Fig. 1. To reduce the chance of missing factors of 2 in coding, Eq. (2) is preferred.

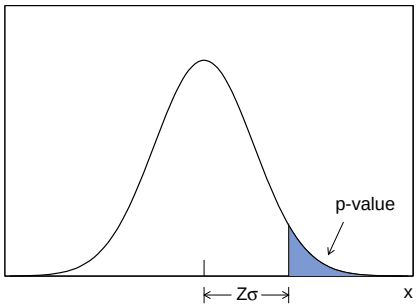

Figure 1: Illustration of the correspondence between the significance Z and a p-value.

Often in HEP, a significance of $Z = 3$ is regarded as "evidence", and $Z = 5$ is taken as "discovery". These correspond to p-values of 1.35×10^{-3} and 2.87×10^{-7}, respectively. Of course to actually draw a firm conclusion about whether a discovery has been made, more than only the p-value should be taken into account. For example, the plausibility of the proposed alternative and the degree to which it describes the data are relevant, as is the reliability of the background model used to determine the p-value and significance.

3 A simple counting experiment

Consider an experiment where one measures a number of events n in a region where signal is expected to be present. Suppose first that the expected number of background events, b, has been determined with negligible uncertainty, and that a signal model predicts an expected number of events s. Thus the number n will follow a Poisson distribution with a mean of $s + b$.

When searching for evidence of a new type of event, we usually regard an observation of a greater number of events as an indication that signal could be present. If we see fewer events than expected from background alone, this is usually not regarded as a discovery of a new phenomenon, but rather an indication of a statistical fluctuation or that the background has been overestimated. Therefore the region of n values that constitute equal or lesser compatibility with the hypothesis of background-only (i.e., $s = 0$) is given by $n \geq n_{\mathrm{obs}}$,

where $n_{\rm obs}$ is the number actually found. The p-value of the background-only hypothesis is therefore

$$p = \sum_{n=n_{\rm obs}}^{\infty} P(n;b) = \sum_{n=n_{\rm obs}}^{\infty} \frac{b^n}{n!} e^{-b} \,, \tag{4}$$

where $P(n;b)$ is the Poisson probability to observe n events for a mean of b. To carry out the sum of Poisson probabilities one can exploit an identity that relates it to the cumulative chi-square distribution F_{χ^2}. This gives

$$\begin{aligned} p &= \sum_{n=n_{\rm obs}}^{\infty} \frac{b^n}{n!} e^{-b} = F_{\chi^2}(2b; 2n_{\rm obs}) \\ &= 1 - \text{TMath::Prob}(2b, 2n_{\rm obs}) \,, \end{aligned} \tag{5}$$

where the final line indicates how to evalute numerically using the root routine `Prob`. Using this with Eq. (2) gives the significance Z. In root code one therefore has

```
Double_t p = 1 - TMath::Prob(2*b, 2*n_obs);
Double_t Z = TMath::NormQuantile(1 - p);
```

This can be called the Poisson significance $Z_{\rm P}$; there are no approximations used in its calculation.

4 Simple counting experiment using the likelihood ratio

The problem above can be reformulated in an equivalent way using the likelihood ratio. The advantage of this approach appears when we generalize to cases with multiple channels or where one must incorporate systematic uncertainties.

We can write the expectation value of the number of events n as

$$E[n] = \mu s + b \,, \tag{6}$$

where μ is a strength parameter defined such that $\mu = 0$ is the background-only hypothesis, and $\mu = 1$ corresponds to background plus the nominal signal. To test a hypothesized value of μ we construct the likelihood ratio

$$\lambda(\mu) = \frac{L(\mu)}{L(\hat{\mu})} \,, \tag{7}$$

where $\hat{\mu}$ is the Maximum Likelihood Estimator (MLE) for μ. Here we require $\hat{\mu} \geq 0$, as this is the usual situation for a physically meaningful signal model. Maximizing the likelihood function with this constraint gives

$$\hat{\mu} = \begin{cases} \frac{n-b}{s} & n \geq b \\ 0 & \text{otherwise} \,, \end{cases} \tag{8}$$

$$\tag{9}$$

In addition to $\lambda(\mu)$ it is convenient to define

$$q_\mu = -2 \ln \lambda(\mu) \ . \tag{10}$$

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, which corresponds to having a small value for q_μ.

Using the ratio of Poisson probabilities for the hypothesis $\mu = 0$ gives

$$q_0 = -2 \ln \lambda(0) = -2n \ln \frac{b}{\hat{\mu} s + b} - 2 \hat{\mu} s \ , \tag{11}$$

where $\hat{\mu}$ is given by Eq.(8). Suppose the data results in a value of $q_0 = q_{0,\mathrm{obs}}$. The level of agreement between the data and the hypothesis $\mu = 0$ is given by the p-value,

$$p = P(q_0 \geq q_{0,\mathrm{obs}} | \mu = 0) \ . \tag{12}$$

Here one always has a larger value of q_0 (less compatibility with the $\mu = 0$ hypothesis), for an increasing number of events found n. That is, one finds $q_0 \geq q_{0,\mathrm{obs}}$ for $n \geq n_{\mathrm{obs}}$. Thus the p-value from Eq. (12) is exactly the same as that from Eq. (4).

For a sufficiently large expected number of events, the statistic q_μ can be treated as a continuous variable, and thus the p-value of a hypothesized value μ can be written

$$p = \int_{q_{\mathrm{obs}}}^{\infty} f(q_\mu | \mu) \, dq_\mu \ , \tag{13}$$

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

Thus 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 for which one gives hypothesized values in the numerator and MLE values in the denominator of the likelihood ratio (39), 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 many cases of interest, the data samples are 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.

Assuming as above only non-negative event rates, the maximum-likelihood estimator for μ is constrained to $\hat{\mu} \geq 0$. If $n_{\mathrm{obs}} < b$, i.e., the observed number of events is below the level predicted by the background alone, then the maximum of the likelihood occurs for $\mu = 0$, and thus the likelihood ratio is

$$\lambda(0) = \frac{L(0)}{L(\hat{\mu})} = 1 \ , \tag{14}$$

In this case the statistic $q_0 = -2 \ln \lambda(0)$ is thus 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

$$f(q_0|0) = w f_{\chi^2_1}(q_0) + (1 - w)\delta(q_0) \; . \tag{15}$$

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

Consider now the variable

$$u = \sqrt{q_0} = \sqrt{-2 \ln \lambda(0)} \; , \tag{16}$$

which has the pdf

$$f(u) = \Theta(u) w \sqrt{\frac{2}{\pi}} e^{-u^2/2} + (1 - w)\delta(u) \; , \tag{17}$$

where $\Theta(u) = 1$ for $u \geq 0$ and is zero otherwise. The second term in (17) 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

$$p = P(u \geq \sqrt{q_0}) = 2w \int_{\sqrt{q_0}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \, du = 2w(1 - \Phi(\sqrt{q_0})) \; . \tag{18}$$

Combining this with equation (2) for the significance Z gives

$$Z = \Phi^{-1}(1 - 2w(1 - \Phi(\sqrt{q_0}))) \; . \tag{19}$$

In the usual case where the weights of the chi-square and delta-function terms are equal, i.e., $w = 1/2$, equation (19) reduces to to the simple formula

$$Z = \sqrt{q_0} = \sqrt{-2 \ln \lambda(0)} \; . \tag{20}$$

For an observed number of events n this becomes

$$Z_{\mathrm{W}} = \sqrt{2n \ln(1 + \hat{\mu} s/b) - 2\hat{\mu} s} \; , \tag{21}$$

where $\hat{\mu}$ is given by Eq. (8) and the subscript W indicates that the approximation relies on the asymptotic chi-square distribution from Wilks' theorem.

5 Sensitivity for a simple counting experiment

To characterize the sensitivity of a planned experiment, one can give the expected (e.g., mean or median) significance assuming a given signal model. For the case of discovery (testing $\mu = 0$, one is most often interested in the expected significance assuming the nominal signal model, i.e., $\mu = 1$. This can be done by generating data by Monte Carlo with $\mu = 1$, and looking at the distribution of q_0.

The median significance can be found by substituting the median value of the number of events n, which is approximately equal to $s + b$ when this is sufficiently large. If this is to be done using the Poisson p-value from Eq. (4), then one needs to choose an integer value to represent the median, e.g., $n_{\rm med} \approx \text{floor}(s + b)$. Then the median significances is

$$\text{med}[Z_{\rm P}] = \Phi^{-1}(1 - F_{\chi^2}(2b; 2n_{\rm med})) \; . \tag{22}$$

Or when using the approximate formula (21) based on the asymptotic distribution of q_0 from Wilks theorem, a good approximation can be found simply by substituting $s + b$ for n (the "Asimov" data value), giving

$$\text{med}[Z_{\rm W}] = \sqrt{2\left((s+b)\ln(1+s/b) - s\right)} \; . \tag{23}$$

If one then takes the limit $s \ll b$, then by expanding the logarithm in (22) and retaining terms up to order s^2, one finds

$$\text{med}[Z_{\rm W}] \approx \frac{s}{\sqrt{b}} \; . \tag{24}$$

Thus in the limit of small s/b, one recovers the widely used formula for Gaussian distributed data.

6 Case with multiple bins and systematic uncertainties

In this section the likelihood ratio method above is extended to cover the case of multiple channels or bins and also where the model includes additional adjustable parameters beyond the parameter of interest μ. Usually these *nuisance parameters* are constrained by including additional measurements, e.g., from control samples that provide information on background rates.

The nuisance parameters correspond to systematic uncertainties in the model. By including enough additional adjustable parameters one can in general improve a model to the point where for at least some point in its parameter space it can be regarded as correct. The price one pays is that the nuisance parameters degrade the significance with which one can reject the background-only hypothesis.

Suppose that the data from an experiment consist of a histogram of some discriminating variable x, which gives values $\mathbf{n} = (n_1, \ldots, n_N)$. In some cases one may consider a histogram with only one bin, i.e., the measured outcome is simply a number of candidate events found. The number of entries in bin i, n_i, is modeled as a Poisson variable with mean value

$$E[n_i] = \mu s_i + b_i \; , \tag{25}$$

where s_i and b_i are the expected number of signal and background events in bin i and as before μ is a strength parameter.

The expected signal and background for bin i can be written

$$s_i = s_{\rm tot} \int_{\text{bin}\,i} f_s(x; \boldsymbol{\theta}_s)\, dx \; , \tag{26}$$

$$b_i = b_{\rm tot} \int_{\text{bin}\,i} f_b(x; \boldsymbol{\theta}_b)\, dx \; , \tag{27}$$

where $s_{\rm tot}$ and $b_{\rm tot}$ are the total expected numbers of events in the histograms, $f_s(x;\boldsymbol{\theta}_s)$ and $f_b(x;\boldsymbol{\theta}_b)$ are the probability density functions (pdfs) of x for signal and background, and $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_b$ represent sets of *shape parameters*.

The parametric forms of the pdfs $f_s(x;\boldsymbol{\theta}_s)$ and $f_b(x;\boldsymbol{\theta}_b)$ are determined from Monte Carlo simulations or data control samples. In the following we will use $\boldsymbol{\theta} = (\boldsymbol{\theta}_s, \boldsymbol{\theta}_b, b_{\rm tot})$ to refer to all of the nuisance parameters. The signal normalization $s_{\rm tot}$ here is not an adjustable parameter, but rather is fixed to the prediction of the nominal signal model.

In addition to the measured histogram $\mathbf{n}$, some search channels also make use of a set of subsidiary measurements $\mathbf{m} = (m_1, \ldots, m_M)$ in control regions where one expects mainly background events. These can be modeled as being Poisson distributed with mean values

$$E[m_i] = u_i(\boldsymbol{\theta}) \ , \tag{28}$$

where the u_i are calculable quantities depending on a set of parameters, at least some of which are the same as those entering into the predictions for s_i and b_i above. In practice the subsidiary measurements are constructed so as to provide information on the background normalization $b_{\rm tot}$ and sometimes also on its shape.

The likelihood function is the product of Poisson probabilities for all bins:

$$L(\mu,\boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \quad \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!}\, e^{-u_k} \ . \tag{29}$$

Equivalently the log-likelihood is

$$\ln L(\mu,\boldsymbol{\theta}) = \sum_{j=1}^{N} \left(n_j \ln(\mu s_j + b_j) - (\mu s_j + b_j)\right) \ + \ \sum_{k=1}^{M} \left(m_k \ln u_k - u_k\right) + C \ , \tag{30}$$

where C represents terms that do not depend on the parameters and thus can be dropped. Here and in (38) the parameters $\boldsymbol{\theta}$ enter through Eqs. (26), (27), and (28).

In the case of several independent search channels, the method described above is generalized in a straightforward manner. For each channel i there is a likelihood function $L_i(\mu,\boldsymbol{\theta}_i)$. Its general form is given by Eq. (38), except that all quantities carry an additional index i to label the channel except the global strength parameter μ, which is assumed to be the same for all channels. Since the channels are statistically independent, the full likelihood function is given by the product

$$L(\mu,\boldsymbol{\theta}) = \prod_i L_i(\mu,\boldsymbol{\theta}_i) \ , \tag{31}$$

where $\boldsymbol{\theta}$ here represents all of the nuisance parameters.

Systematic uncertainties are effectively included in the analysis through the nuisance parameters $\boldsymbol{\theta}$. The model must be sufficiently flexible, i.e., it must contain enough parameters, so that for at least some point in its parameter space it can be regarded as representing the truth. One must exercise some restraint in achieving this, however, as an increasing number of nuisance parameters leads to a decrease in sensitivity to the parameters of interest. Some of the components of $\boldsymbol{\theta}$ may be common among different channels, e.g., parameters relating to uncertainty in the integrated luminosity. These then represent a common (correlated) systematic uncertainty.

As an example, consider the signal efficiency ε that enters in the relation between the cross section and expected number of signal events. Suppose the efficiency has been estimated to have a value $\hat{\varepsilon}$ and systematic uncertainty $\sigma_{\hat{\varepsilon}}$. To incorporate this uncertainty into the model, we can regard the measured value $\hat{\varepsilon}$ as a random variable whose true value ε is treated as a nuisance parameter. For the pdf $f_\varepsilon(\hat{\varepsilon};\varepsilon,\sigma_{\hat{\varepsilon}})$ one could use, e.g., a Gaussian distribution centred about ε, or for a quantity such as the efficiency which must lie in the range $0 \leq \varepsilon \leq 1$ one could use a pdf that automatically satisfies this constraint (e.g., a beta distribution). For whatever choice is deemed appropriate, the likelihood (38) is multiplied by $f_\varepsilon(\hat{\varepsilon};\varepsilon,\sigma_{\hat{\varepsilon}})$, evaluated with the best estimate $\hat{\varepsilon}$, and the parameter ε is included in the set of nuisance parameters $\boldsymbol{\theta}$.

To test a hypothesized value of μ we construct the profile likelihood ratio,

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} \; . \tag{32}$$

Here $\hat{\hat{\boldsymbol{\theta}}}$ in the numerator denotes the value of $\boldsymbol{\theta}$ that maximizes L for the specified μ, i.e., it is the conditional maximum-likelihood estimator (MLE) of $\boldsymbol{\theta}$ (and thus is a function of μ). The denominator is the maximized (full) likelihood function, i.e., $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ are the MLEs. The presence of the nuisance parameters broadens the profile likelihood ratio as a function of μ relative to what one would have if their values were fixed. This reflects the loss of information about μ due to the systematic uncertainties.

Although the situation now appears substantially more complicated than what was encountered in the previous sections, we can exploit the fact that the distribution assuming μ of $-2\ln\lambda(\mu)$, even with multiple channels and nuisance parameters, still approaches a chi-square pdf for one degree of freedom. The validity of this for a finite data sample should be checked, e.g., with Monte Carlo, but in practice the approximation is found to be quite good even for small data samples. Providing this is true, we can obtain the discovery significance in the same manner as above, namely,

$$Z = \sqrt{q_0} = \sqrt{-2\ln\lambda(0)} \; . \tag{33}$$

To estimate the median discovery significance assuming the nominal signal model, i.e., $\mu = 1$, in many cases it is a good approximation to simply evaluate Z with $n_i = s_i + b_i$ (the Asimov data set).

7 Simple counting experiment with background uncertainty

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

$$E[n] = \mu s + b_{\rm tot} \; , \tag{34}$$

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.,

$$b_{\rm tot} = \sum_{i=1}^{N} b_i \ . \tag{35}$$

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

$$E[m_i] = \tau_i b_i \ . \tag{36}$$

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,

$$\tau_i = \frac{L_{{\rm MC},i}}{L_{\rm data}} \ . \tag{37}$$

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:

$$L(\mu, \mathbf{b}) = \frac{(\mu s + b_{\rm tot})^n}{n!} e^{-(\mu s + b_{\rm tot})} \prod_{i=1}^{N} \frac{(\tau_i b_i)^{m_i}}{m_i!} e^{-\tau_i b_i} \ . \tag{38}$$

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

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\mathbf{b}}})}{L(\hat{\mu}, \hat{\mathbf{b}})} \tag{39}$$

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.

To calculate the value of $\lambda(\mu)$ 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. 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 $L = 1\,{\rm fb}^{-1}$ and the equivalent luminosity of the MC samples is shown Table 1.

For $L = 1\,{\rm fb}^{-1}$, $s = 312$ 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 ${\rm pb}^{-1}$.

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

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 8.

8 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

$$\hat{\mu} = \begin{cases} \frac{n - m/\tau}{s} & n \geq m/\tau \\ 0 & \text{otherwise} , \end{cases} \tag{40}$$

$$\hat{b} = \begin{cases} m/\tau & n \geq m/\tau \\ \frac{n+m}{\tau+1} & \text{otherwise} , \end{cases} \tag{41}$$

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

$$\hat{\hat{b}} = \frac{n+m}{\tau+1} . \tag{42}$$

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

$$\ln \lambda(0) = \begin{cases} \psi(m, \tau\hat{\hat{b}}) + \psi(n, \hat{\hat{b}}) - \psi(m, \tau\hat{b}) - \psi(n, \hat{\mu}s + \hat{b}) & n \geq m/\tau , \\ 0 & \text{otherwise} , \end{cases} \tag{43}$$

where

$$\psi(x, y) = \begin{cases} 0 & x = y = 0 \ , \\ x \ln y - y & \text{otherwise} \ . \end{cases} \tag{44}$$

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

$$q_0 = -2 \ln \lambda(0) = 2s \ln(1 + \tau) = 28.5 \ . \tag{45}$$

Using the asymptotic formula (20) for the significance gives

$$Z = \sqrt{q_0} = 5.3 \ . \tag{46}$$

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 the formulae for the significance revert to those found in Section 3 for the case with known background.

Figure 2 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 valid for large τ, and the limiting formula valid for large τ and $s \ll b$.

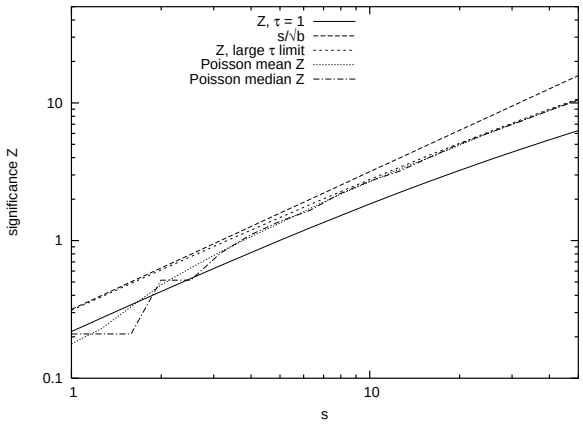

Figure 2: The significance Z as a function of the expected signal s according to several formulae (see text).

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. 2 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 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 3 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.

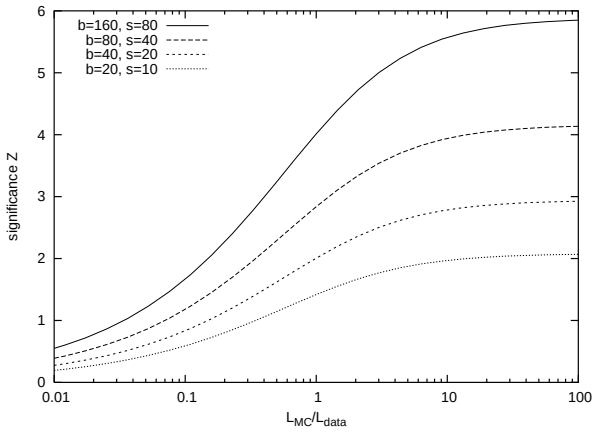

Figure 3: The significance Z as a function of the luminosity ratio $\tau = L_{\rm MC}/L_{\rm data}$ for several values of b and s.

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.