DRAFT 1.2
Glen Cowan
12 May, 2011

Marginal vs. Profile Likelihood for systematic uncertainties

This note describes treatment of systematic uncertainties using a Bayesian averaged (i.e., marginal) likelihood as opposed to the profile likelihood. The problem of uncertain background in a Poisson counting experiment is addressed in Sec. 1 for a Gamma prior, which corresponds to having a Poisson measurement to constrain the background. In Sec. 2 the same problem is treated using a log-normal prior, which corresponds to having a control measurement that also follows a log-normal distribution.

1 Uncertainty on background in Poisson mean

As a simple example, suppose one has an experiment that counts n events, modeled as following a Poisson distribution with mean $\mu s + b$, where s is the nominal signal rate, b is the expected background rate, and μ is a strength parameter. In this example we regard s as known. The probability for n is therefore

$$P(n|\mu, b) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \; . \tag{1}$$

Suppose b is not known exactly. There are now two basic approaches to the problem, depending on how we choose to quantify this uncertainty. The approach of Bayesian model averaging (marginalization) is described in Sec. 1.1, and the method of profiling is shown in Sec. 1.2.

1.1 Marginalization approach for Poisson mean with gamma prior

In the marginalization approach to the problem, one would say that the uncertainty in b is characterized by some Bayesian prior pdf $\pi(b)$. The model for n is taken as the Bayesian model average,

$$P_{\rm m}(n|\mu) = \int P(n|\mu, b)\pi(b)\, db \; , \tag{2}$$

i.e., the nuisance parameter is eliminated by marginalization, hence the subscript m. The probability (2) is thus no longer Poisson but rather some other (still discrete) law, which will in general be broader than Poisson. Equation (2) could be used to construct a test statistic, e.g.,

$$q_{\rm m} = -2 \ln \frac{P_{\rm m}(n|1)}{P_{\rm m}(n|0)} \; . \tag{3}$$

The sampling distribution of this statistic is found by generating n according to Eq. (2), and using this to evaluate $q_{\rm m}$. In practice, the integral in (2) would be carried out by sampling b from $\pi(b)$, and then using that b in Eq. (1) to generate n.

A difficulty with the statistic $q_{\rm m}$ as defined in Eq. (3) is that one must be able to evaluate the functions $P_{\rm m}(n|1)$ and $P_{\rm m}(n|0)$ for arbitrary n. But after marginalization, this is no longer a simple Poisson distribution, but rather more complicated. Thus one would first have to determine $P_{\rm m}(n|1)$ and $P_{\rm m}(n|0)$, e.g., using Monte Carlo, and store the result so that it could be used to evaluate $q_{\rm m}$.

Alternatively, one could generate the n according to the marginalized model, but then use it to evaluate a test statistic defined using the original Poisson probability from Eq. (1) for n, i.e.,

$$q'_{\rm m} = -2 \ln \frac{P(n|1,b)}{P(n|0,b)} = -2 \left(n \ln \left(\frac{s+b}{b} \right) - s \right) . \tag{4}$$

For this one requires an assumed value of b, e.g., the mean of $\pi(b)$. For the present study this is the approach followed.

It is important to note that although the sampling distribution of n is found using a Bayesian argument, the subsequent analysis follows a frequentist approach. Alternatively one could assume a prior for μ (or more generally, a joint prior for μ and b), and use this to find the Bayesian posterior pdf for μ. This 'fully Bayesian' approach is not considered here.

1.2 Profile approach for Poisson mean with Poisson constraint

In the profile-likelihood approach, one treats b as a free (nuisance) parameter. If one has no other measurement or constraint on b, then inference about μ is impossible. Any value of μ will fit the data if one is allowed to adjust b arbitrarily. So to proceed we must have some measurement that constrains b. Suppose this is a control measurement of a number of events m, assumed to follow a Poisson distribution with mean τb, where we take τ to be a known scale factor, i.e.,

$$P(m|b) = \frac{(\tau b)^m}{m!} e^{-\tau b} . \tag{5}$$

The full experimental outcome is now regarded as consisting of n (the main measurement) and m (the control measurement). Assuming n and m are independent, the full model is simply the product of the two Poisson probabilities:

$$P(n,m|\mu,b) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \, \frac{(\tau b)^m}{m!} e^{-\tau b} . \tag{6}$$

One can construct a test statistic analogous to the one used above in the Bayesian approach by using the ratio of profile likelihoods, i.e.,

$$q_{\rm p} = -2 \ln \frac{P(n,m|1,\hat{\hat{b}}(1))}{P(n,m|0,\hat{\hat{b}}(0))} . \tag{7}$$

Here the arguments 1 and 0 in the numerator and denominator refer to the hypothesized values of μ, and $\hat{\hat{b}}(\mu)$ is the value of b that maximizes the likelihood (6) for the given value of μ.

1.3 Comparison of marginalization and profile approaches

One may ask whether there is some prior $\pi(b)$ such that the distribution of the statistics based on marginalization and profiling, q_{m} and q_{p} will lead to identical tests. Here one must keep in mind that to generate a value of q_{m}, one samples n according to Eq. (1), which is equivalent to generating b from $\pi(b)$, and using this b to sample n from the Poisson distribution (1). In contrast, to generate q_{p} one fixes b to some point (this could be related to the mean of $\pi(b)$), and then one generates n and m according to Eqs. (1) and (5), respectively. Thus the value of q_{p} is a function of two independent, discrete random variables, whereas q_{m} only depends on the value n, so effects due to discreteness do not enter in the same way for the two statistics.

But beyond issues related to discreteness, one can find what the Poisson measurement of m would imply about b, and use this information to determine a prior $\pi(b)$. Suppose the original prior for b, i.e., before even the control measurement is carried out, is $\pi_0(b)$. This may be called the 'ur-prior', using the German prefix meaning original or primordial. Then Bayes' theorem gives the posterior probability

$$p(b|m) \propto P(m|b)\pi_0(b) \; . \tag{8}$$

If we take the ur-prior to be constant, then the probability for b becomes

$$p(b|m) \propto \frac{(\tau b)^{-m}}{m!} e^{-\tau b} \; , \tag{9}$$

which has the general form of a gamma distribution for b. Normalizing this to unit area over $0 \le b < \infty$ gives

$$p(b|m) = \frac{\tau^{m+1} b^m e^{-\tau b}}{m!} \; , \tag{10}$$

which has a mean of $(m+1)/\tau$ and variance of $(m+1)/\tau^2$.

One may now say that this information about b, i.e., after measurement of m but before the measurement of n, was what led to the prior $\pi(b)$ used in the approach of Sec. 1.1 to find the marginal likelihood. That is, this prior is the same as the pdf of b given m,

$$\pi(b) = p(b|m) = \frac{\tau^{m+1} b^m e^{-\tau b}}{m!} \; . \tag{11}$$

It is important to note that the correspondence between a Poisson distributed measurement m and the gamma prior depends on taking a constant for the ur-prior $\pi_0(b)$.

As an example, consider $s = 10$, $\tau = 1$ and a Poisson distributed control measurement for the background that yielded $m = 20$ events. The Maximum Likelihood estimator (MLE) for b is thus $\hat{b} = m/\tau = 20$. For computing the distribution of the statistic q_{p}, in this study, therefore, the parameter values $s = 10$, $b = 20$ were used.

In a real analysis, however, one would have actual data values, n and m, and on the basis of these one can compute the conditional MLE $\hat{\hat{b}}(\mu)$, i.e., the value of b that maximizes the likelihood for the given value of μ. One would then use $\hat{\hat{b}}(0)$ to generate the data needed to determine the distribution $f(q_{\mathrm{p}}|0)$, and $\hat{\hat{b}}(1)$ to compute $f(q_{\mathrm{p}}|1)$.

For marginalization, the corresponding prior should be a gamma pdf with mean $E[b] = (m+1)/\tau = 21$ and standard deviation $\sigma = \sqrt{m+1}/\tau = \sqrt{21} = 4.58$. Figure 1 shows the resulting distributions for the statistics $q'_{\rm m}$ and $q_{\rm p}$. The different effect of discreteness is clearly visible; beyond that their discriminating power is similar.

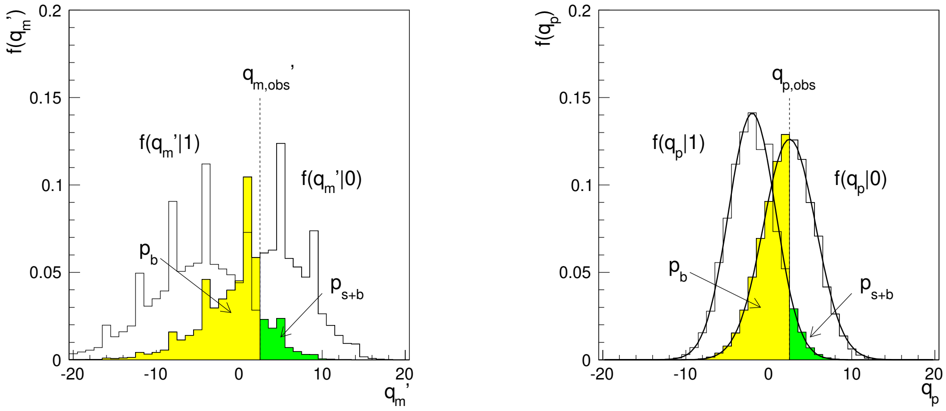

Figure 1: Distributions of the test statistics (a) $q'_{\rm m}$ and (b) $q_{\rm p}$ assuming both the background-only and signal-plus-background hypotheses. An indicative observed value of the statistics shows how one would obtain p-values for the two hypotheses. For $q_{\rm p}$, the solid curves show the asymptotic distributions from Ref. [1].

2 Uncertainty modeled with a log-normal pdf

In this section, the uncertainty on the mean number of background events, b, is modeled with a log-normal prior (see, e.g., [2]), i.e.,

$$\pi_b(b) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{b} \exp\left[-\frac{(\ln(b/b_0))^2}{2\sigma^2} \right] . \tag{12}$$

This corresponds to having a Gaussian distribution,

$$\pi_\beta(\beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\beta - \beta_0)^2}{2\sigma^2} \right] , \tag{13}$$

for

$$\beta = \ln b , \tag{14}$$

with mean value $\beta_0 = \ln b_0$ and standard deviation σ, which in the following we write as σ_β.

Following Ref. [2], we define $\kappa = e^{\sigma_\beta}$ and identify $\sigma_{\rm rel} = \kappa - 1 = e^{\sigma_\beta} - 1$ as the relative uncertainty on b. That is, a 20% background uncertainty corresponds to taking $\sigma_{\rm rel} = 0.2$ and then using

$$\sigma_\beta = \ln(1 + \sigma_{\rm rel}) . \tag{15}$$

The Bayesian-averaged model can be found in the same manner as was done in Sec. 1.1 using the gamma prior. That is, the probability for n given μ is still given by Eq. (2), but now using the log-normal prior (12). In practice one generates values of n by sampling β from the Gaussian distribution (13), then using $b = e^{\beta}$ as the expected background in the Poisson distribution (1) to generate n.

As pointed out at the ATLAS/CMS Higgs Combination Meeting (5.5.11), use of the log-normal prior corresponds to having a measurement $b_{\rm meas}$ of b with a log-normal sampling distribution, or equivalently to having a Gaussian measurement $\beta_{\rm meas}$ of the parameter $\beta = \ln b$, i.e.,

$$p(\beta_{\rm meas}|\beta) = \frac{1}{\sqrt{2\pi}\sigma_{\beta}} e^{-(\beta_{\rm meas}-\beta)/2\sigma_{\beta}^2} \; . \tag{16}$$

Given a value of $\beta_{\rm meas}$, one would find the posterior pdf for β,

$$p(\beta|\beta_{\rm meas}) \propto p(\beta_{\rm meas}|\beta)\pi_{0,\beta}(\beta) \; , \tag{17}$$

where $\pi_{0,\beta}(\beta)$ is the Bayesian ur-prior of β. If we take this as a constant, then this implies an ur-prior for b of

$$\pi_{0,b}(b) = \pi_{0,\beta}(\beta) \left| \frac{d\beta}{db} \right| \propto \frac{1}{b} \; . \tag{18}$$

Assuming a constant $\pi_{0,\beta}(\beta)$, Eq. (17) for the posterior probability of β is a Gaussian distribution centred about $\beta_{\rm meas}$. In the Bayesian approach one would then use this as the prior for β, which is based in the control measurement $\beta_{\rm meas}$ and the constant ur-prior. Thus having a log-normal distributed measurement for b in the profile approach corresponds to using a log-normal prior for the parameter b in the marginalization approach.

To implement such a constraint in the profile approach, one can assume that the data consist of the main measurement n, which follows a Poisson distribution with mean $\mu s + e^{\beta}$, and a control measurement $\beta_{\rm meas}$, which is Gaussian distributed about β with standard deviation σ_{β}.

Distributions of the test statistics $q_{\rm m}'$ and $q_{\rm p}$ assuming the background-only ($\mu = 0$) and signal-plus-background ($\mu = 1$) hypotheses are shown in Fig. 2. The parameters used were $s = 10$, $b = 20$ and a relative uncertainty in b of $\sigma_{\rm rel} = 0.2$.

Distributions of the test statistics q_0 and q_1 as defined in Ref. [1] are shown with the asymptotic (large-sample) distributions in Fig. 3. These are based on the profile likelihood ratio,

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} \; . \tag{19}$$

where $\boldsymbol{\theta}$ represents any nuisance parameters (in the present problem, b), and as before a single hat indicates the MLE and a double hat indicates the conditional MLE for the specified value of μ.

The statistic q_0 would be used to test the background-only ($\mu = 0$) hypothesis against an alternative with $\mu > 0$. It is defined as

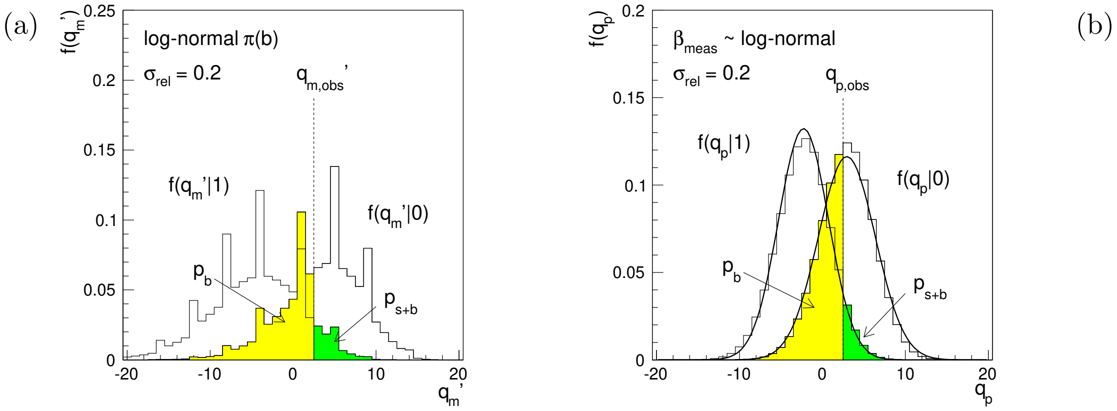

Figure 2: Distributions of the test statistics (a) q'_{m} and (b) q_{p} assuming both the background-only and signal-plus-background hypotheses using the log-normal of Sec. 2. For q_{p}, the solid curves show the asymptotic distributions from Ref. [1].

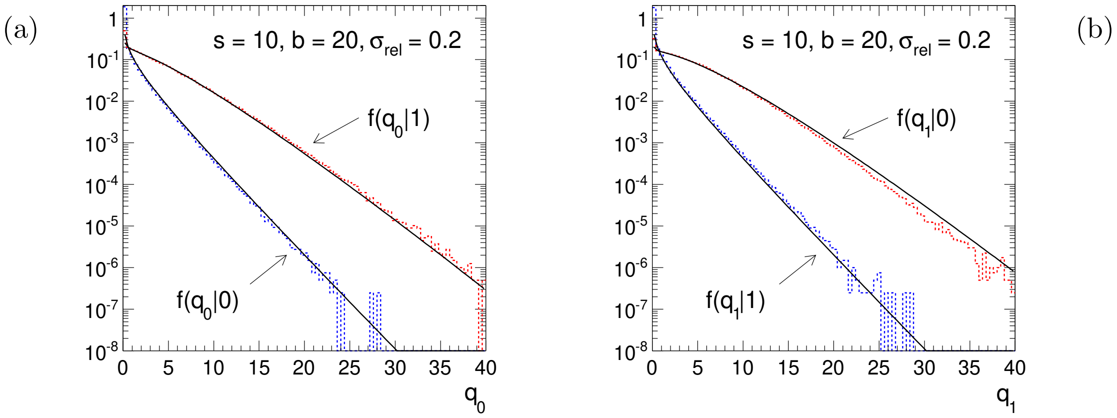

Figure 3: Distributions of the test statistics (a) q_0 and (b) q_1 assuming both the background-only and signal-plus-background hypotheses using the log-normal of Sec. 2. The solid curves show the asymptotic distributions from Ref. [1].

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \ , \\ 0 & \hat{\mu} < 0 \ , \end{cases} \tag{20}$$

where $\lambda(0)$ is the profile likelihood ratio for $\mu = 0$. For purposes of setting an upper limit, one tests values of μ against the alternative consisting of lower values. For this test we use the statistic

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \leq \mu \ , \\ 0 & \hat{\mu} > \mu \ . \end{cases} \tag{21}$$

The statistic q_1 is the special case of q_μ for $\mu = 1$. Note, however, that q_0 has a separate definition and is not a special case of q_μ. Both are defined so as to give one-sided tests of a hypothesized value of μ.

As can be seen from Eq. (19), the quantity $\lambda(\mu)$ satisfies $0 \leq \lambda(\mu) \leq 1$, with a value of $\lambda(\mu)$ closer to unity reflecting reflecting a higher level of agreement between data and the hypothesized value of μ. Thus higher values of q_0 or q_μ reflect increasing incompatibility between the hypothesized value of μ and the data. The p-value of the hypothesized μ can therefore be found from

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu \ . \tag{22}$$

The upper limit on μ at confidence level $1-\alpha$ is the highest value of μ not rejected in a test of size α. In practice this is found by setting $p_\mu = \alpha$ and solving for μ. With this procedure, a strong downward fluctuation of the data can result in an upper limit that is substantially smaller than the intrinsic resolution of the measurement, even to the point where all values of μ are excluded. To protect against this, one can base the upper limit not on the usual p-value but on the quantity CL_s [3], defined as

$$\mathrm{CL}_s = \frac{p_\mu}{1 - p_0} \ . \tag{23}$$

Alternatively, one may determine the Power-Constrained Limit [4]. Here one regards a value of μ as excluded if two criteria are satisfied, namely, it is excluded by the usual test ($p_\mu < \alpha$) and in addition one has sufficient sensitivity to μ. The measure of sensitivity adopted by ATLAS is the power of a test of μ with respect to the background-only alternative, which is required to be greater than a minimum threshold M_{min}:

$$P(p_\mu < \alpha|0) \geq M_{\mathrm{min}} \ . \tag{24}$$

The choice of M_{min} is a matter of convention. The value of $M_{\mathrm{min}} = \Phi(-1) = 0.1587$ has been used in recent ATLAS analyses.

The p-value for discovery may be converted into the equivalent Gaussian significance Z by the relation (see, e.g., [1]),

$$Z = \Phi^{-1}(1 - p) \ , \tag{25}$$

where Φ^{-1} is the standard normal quantile (inverse of the standard normal cumulative distribution). Requiring $Z > 5$ (a 5σ effect) corresponds to $p < 2.9 \times 10^{-7}$.

To determine the p-value for a hypothesized value of μ one requires the distribution $f(q_\mu|\mu)$. This may be found either from Monte Carlo or by using the asymptotic formulae discussed in Ref. [1]. The asymptotic distributions $f(q_0|0)$ and $f(q_1|1)$ are both given by a delta function at zero with a weight of one half plus a chi-square pdf for one degree of freedom, also with weight one half. These pdfs are thus asymptotically independent of all nuisance parameters.

Using the results from Ref. [1], the asymptotic approximation for the p-value for a hypothesized value of μ is has the simple formula

$$p_\mu = 1 - \Phi(\sqrt{q_\mu}) \;, \tag{26}$$

where Φ is the standard normal cumulative distribution. Similarly, for the case of testing $\mu = 0$ for discovery, the formula for the p-value has the same form as (26), or equivalently one can write the discovery significance as

$$Z = \sqrt{q_0} \;. \tag{27}$$

Although the asymptotic formulae are exact only in the large-sample limit, Monte Carlo studies (see, e.g., [1]) show that they are reasonably accurate even for relatively small data samples. Even in those cases where one is not prepared to trust the large-sample formulae, use of the asymptotic expressions provides an important check of the procedure.

References

[1] G. Cowan, K. Cranmer, E. Gross and O. Vitells, Eur. Phys. J. C (2011) 71:1554; arXiv:1007.1727 [physics.data-an].

[2] R.D. Cousins, A. Korytov et al., (CMS Collaboration) note on log-normal prior (need reference).

[3] T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

[4] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Power-Constrained Limits* (in preparation; draft paper available at `www.pp.rhul.ac.uk/ cowan/stat/pcl/pcl.pdf`).