PROCEEDINGS OF SCIENCE

Tackling limited simulation and small signals

Carlos A. Argüelles
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
E-mail: caad@mit.edu

Austin Schneider*
Dept. of Physics and Wisconsin IceCube Particle Astrophysics Center, University of Wisconsin, Madison WI 53706, USA
E-mail: aschneider@icecube.wisc.edu

Tianlu Yuan
Dept. of Physics and Wisconsin IceCube Particle Astrophysics Center, University of Wisconsin, Madison WI 53706, USA
E-mail: tyuan@icecube.wisc.edu

We present a new, analytic, Poisson likelihood derived, technique to account for the statistical uncertainties inherent in simulation samples of limited size. This method has better coverage properties than other techniques, is valid for small data samples, and maintains good computational performance.

36th International Cosmic Ray Conference -ICRC2019-
July 24th - August 1st, 2019
Madison, WI, U.S.A.

*Speaker.

PoS(ICRC2019)1005

For binned data, the Poisson likelihood is taken to be the probability to observe events in a bin and is commonly used in high-energy physics and particle astrophysics experiments. Given an exact expectation rate, $\lambda(\vec{\theta})$, the probability of observing an integer k events is

$$\mathcal{L}(\tilde{\theta}|\mathrm{k}) = \mathrm{Poisson}(\mathrm{k};\lambda(\tilde{\theta})) = \frac{\lambda(\tilde{\theta})^{\mathrm{k}}\mathrm{e}^{-\lambda(\tilde{\theta})}}{\mathrm{k!}}, \tag{1}$$

where $\vec{\theta}$ is some set of physics parameters that determine λ. Given the stochastic nature of processes in particle physics, exactly determining λ is often not possible and requires Monte Carlo (MC) simulations. Simulation is often expensive, and so reweighting is employed as it enables a single simulation to be used to describe many physical hypotheses [1].

In such scenarios, an *ad hoc* likelihood is commonly used, where λ is simply taken to be the sum of the weights in each bin, namely

$$\mathcal{L}_{\mathrm{AdHoc}}(\vec{\theta}|k) = \frac{\left(\sum_i w_i(\vec{\theta})\right)^k e^{-\left(\sum_i w_i(\vec{\theta})\right)}}{k!}. \tag{2}$$

A notable downside of this *ad hoc* likelihood is that it neglects the statistical uncertainty inherent in estimating $\lambda(\vec{\theta})$ from a simulation of limited size. For expensive simulations, or physical hypotheses far from the original simulation, this uncertainty can be non-negligible [2, 3, 4, 5]. One can account for this uncertainty by incorporating a simulation derived prior on λ, denoted as $\mathcal{P}\left(\lambda|\vec{w}(\vec{\theta})\right)$, that has non-zero variance. Thus, we can write the likelihood as the marginalization of the Poisson likelihood with the prior,

$$\mathcal{L}_{\mathrm{General}}(\vec{\theta}|k) = \int_0^\infty \frac{\lambda^k e^{-\lambda}}{k!}\mathcal{P}\left(\lambda|\vec{w}(\vec{\theta})\right)\, d\lambda. \tag{3}$$

We construct this prior based on the likelihood function of the simulation outcome and a prior on λ, $\mathcal{P}(\lambda)$,

$$\mathcal{P}\left(\lambda|\vec{w}(\vec{\theta})\right) = \frac{\mathcal{L}(\lambda|\vec{w}(\vec{\theta}))\mathcal{P}(\lambda)}{\int_0^\infty \mathcal{L}(\lambda'|\vec{w}(\vec{\theta}))\mathcal{P}(\lambda')\, d\lambda'}, \tag{4}$$

where in our implementation we have chosen we chosen $\mathcal{P}(\lambda)$ to be uniform.

Let us first consider the case where all simulation events in the bin have equal weight. In this scenario we can relate the number of events, m, and the weight of the events, w, to the quantities μ and σ, which are defined

$$\mu \equiv \sum_{i=1}^{m} w_i \text{ and } \sigma^2 \equiv \sum_{i=1}^{m} w_i^2, \tag{5}$$

and satisfy the relationships

$$\mu = wm,\ \sigma^2 = w^2 m,\ w = \sigma^2/\mu, \text{ and } \ m = \mu^2/\sigma^2. \tag{6}$$

The probability of obtaining m events in the simulation bin can be modelled with the Poisson distribution assuming the true but unknown mean $\bar{m}$:

$$\mathrm{Poisson}(M = m;\bar{m}) = \frac{e^{-\bar{m}}\bar{m}^m}{m!}. \tag{7}$$

This allows us to rewrite the likelihood of λ in terms of μ and σ as

$$\mathcal{L}(\lambda|\vec{w}(\vec{\theta})) = \mathcal{L}(\lambda|\mu,\sigma) = \frac{e^{-\lambda\mu/\sigma^2}\left(\lambda\mu/\sigma^2\right)^{\mu^2/\sigma^2}}{(\mu^2/\sigma^2)!}. \tag{8}$$

If the simulation event weights are not all equal, as is usually the case, then we can replace w and m with their "effective" counterparts w_{Eff} and m_{Eff}. These then relate to μ and σ as

$$\mu = w_{\text{Eff}} m_{\text{Eff}} \text{ and } \sigma^2 = w_{\text{Eff}}^2 m_{\text{Eff}}. \tag{9}$$

The replacement redefines the likelihood of our simulation outcome

$$\mathcal{L}(\bar{m}|m_{\text{Eff}}) = \frac{e^{-\bar{m}}\bar{m}^{m_{\text{Eff}}}}{\Gamma(m_{\text{Eff}}+1)}, \tag{10}$$

which, assuming $\lambda = w_{\text{Eff}}\bar{m}$, can be rewritten as

$$\mathcal{L}(\lambda|\vec{w}(\vec{\theta})) = \mathcal{L}(\lambda|\mu,\sigma) = \frac{e^{-\lambda\mu/\sigma^2}\left(\lambda\mu/\sigma^2\right)^{\mu^2/\sigma^2}}{\Gamma(\mu^2/\sigma^2+1)}. \tag{11}$$

To simplify the notation, define

$$\alpha \equiv \frac{\mu^2}{\sigma^2} + 1 \text{ and } \beta \equiv \frac{\mu}{\sigma^2}. \tag{12}$$

Substituting Eq. (11) into Eq. (4) and assuming a uniform $\mathcal{P}(\lambda)$, we obtain

$$\begin{aligned}\mathcal{P}(\lambda|\vec{w}(\vec{\theta})) &= \beta\frac{e^{-\lambda\beta}(\lambda\beta)^{\alpha-1}}{\Gamma(\alpha)} \\ &= \frac{e^{-\lambda\beta}\lambda^{\alpha-1}\beta^{\alpha}}{\Gamma(\alpha)} \\ &= \mathcal{G}(\lambda;\alpha,\beta),\end{aligned} \tag{13}$$

where $\mathcal{G}(\lambda;\alpha,\beta)$ is the gamma distribution, with shape and inverse rate parameters α and β. Finally, this can be substituted for $\mathcal{P}\left(\lambda|\vec{w}(\vec{\theta})\right)$ in Eq. (3) so that

$$\mathcal{L}_{\text{Eff}}(\vec{\theta}|k) = \int_0^\infty \frac{\lambda^k e^{-\lambda}}{k!}\mathcal{G}(\lambda;\alpha,\beta)\,d\lambda \tag{14}$$

$$= \frac{\beta^\alpha\Gamma(k+\alpha)}{k!\,(1+\beta)^{k+\alpha}\,\Gamma(\alpha)} \tag{15}$$

$$= \left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{\sigma^2}+1}\Gamma\left(k+\frac{\mu^2}{\sigma^2}+1\right)\left[k!\left(1+\frac{\mu}{\sigma^2}\right)^{k+\frac{\mu^2}{\sigma^2}+1}\Gamma\left(\frac{\mu^2}{\sigma^2}+1\right)\right]^{-1}. \tag{16}$$

Equation (16) is an effective likelihood, motivated by Poisson statistics, and derived with a Bayesian approach. It incorporates statistical uncertainties inherent in the MC approximation of the rate by encoding the distribution of weights in terms of μ and σ^2. The effective likelihood $\mathcal{L}_{\text{Eff}}$ can be easily substituted for $\mathcal{L}_{\text{AdHoc}}$. A more thorough exposition, along with a generalization for different priors, $\mathcal{P}(\lambda)$, is given in [6].

References

[1] J. S. Gainer, J. Lykken, K. T. Matchev, S. Mrenna, and M. Park, *JHEP* **10** (2014) 078.

[2] R. J. Barlow and C. Beeston, *Comput. Phys. Commun.* **77** (1993) 219–228.

[3] G. Bohm and G. Zech, *Nucl. Instrum. Meth.* **A748** (2014) 1–6.

[4] D. Chirkin, arXiv:1304.0735.

[5] T. Glüsenkamp, *Eur. Phys. J. Plus* **133** (2018) 218.

[6] C. A. Argüelles, A. Schneider, and T. Yuan, *JHEP* **06** (2019) 030.