Likelihood description for comparing data to simulation of limited statistics

Dmitry Chirkin (dima@icecube.wisc.edu) University of Wisconsin, Madison, U.S.A.

Introduction

It is often not possible to construct a probability density function that describes the data. This can happen if there is no analytic description, and the number of parameters is too large so that it is impossible to simulate and tabulate all combinations. In these situations it is still interesting to rank simulation sets performed with different parameters in how well they compare to data. We propose a solution that appears to be better suited to this task than some of the obvious alternatives.

Likelihood description with statistical uncertainties only

Consider a repeatable experiment that is performed n_d times to collect a total of d counts with a per-event expectation of μ_d (we call a single instance of this experiment an "event"). We predict the result of the experiment with the simulation, which collects s counts in n_s simulated events and a per-event expectation of μ_s.
Given that the total count in the combined set of simulation and data is $s+d$, the conditional probability distribution function of observing s simulation and d data counts is

$$P(\mu_s,\mu_d;s,d|s+d)=\frac{(s+d)!}{s!\cdot d!}\cdot\left(\frac{n_s\mu_s}{s+d}\right)^s\cdot\left(\frac{n_d\mu_d}{s+d}\right)^d.$$

An obvious constraint that is implied here is $n_s\mu_s+n_d\mu_d=s+d$, which can be derived from the normalization requirement

$$\sum_{s,d}P(\mu_s,\mu_d;s,d|s+d)=\left(\frac{n_s\mu_s}{s+d}+\frac{n_d\mu_d}{s+d}\right)^{s+d}=1.$$

If the data and simulation are completely unrelated the best possible estimates of μ_s and μ_d are determined by maximization of the probability function given above with the constraint $n_s\mu_s+n_d\mu_d=s+d$, which results in the estimates

$$\mu_s=\frac{s}{n_s},\quad \mu_d=\frac{d}{n_d}.$$

Now, the alternative hypothesis that we could try to verify is that data and simulation counts are described by the same process, i.e., come with the same mean $\mu=\mu_s=\mu_d$. This identity together with the constraint $n_s\mu_s+n_d\mu_d=s+d$ uniquely determines the values of

$$\mu=\mu_s=\mu_d=\frac{s+d}{n_s+n_d}.$$

We can now compare the probabilities of the two of the above hypotheses by forming a likelihood ratio

$$\frac{P(\text{same process})}{P(\text{independent processes})}=\left(\frac{\mu}{s/n_s}\right)^s\cdot\left(\frac{\mu}{d/n_d}\right)^d.$$

This is the expression that we propose for comparison of different simulation sets with data. It can obviously also be derived starting with the Poisson probability

$$P(\mu_s,\mu_d;s,d)=\frac{(n_s\mu_s)^s e^{-n_s\mu_s}}{s!}\cdot\frac{(n_d\mu_d)^d e^{-n_d\mu_d}}{d!}.$$

Adding model errors

The error in describing data with simulation (i.e., describing μ_d with μ_s) is often non-zero. In such a case one may quantify the amount of disagreement between data and simulation with a χ^2:

$$\chi^2=\frac{(\ln\mu_d-\ln\mu_s)^2}{\sigma^2}.$$

Instead of setting $\mu_s=\mu_d$ as in the previous sections we assume that a difference between μ_s and μ_d can exist due to this systematic error and is modeled with a likelihood penalty term

$$\exp\frac{\ln^2(\mu_d/\mu_s)}{-2\sigma^2}.$$

The likelihood ratio is therefore determined as

$$\frac{P(\text{same process})}{P(\text{independent processes})}=\left(\frac{\mu_s}{s/n_s}\right)^s\cdot\left(\frac{\mu_d}{d/n_d}\right)^d\cdot\exp\frac{\ln^2(\mu_d/\mu_s)}{-2\sigma^2},$$

where the μ_s and μ_d are determined by maximizing

$$P(\mu_s,\mu_d;s,d|s+d)=\frac{(s+d)!}{s!\cdot d!}\cdot\left(\frac{n_s\mu_s}{s+d}\right)^s\cdot\left(\frac{n_d\mu_d}{s+d}\right)^d\cdot\exp\frac{\ln^2(\mu_d/\mu_s)}{-2\sigma^2}$$

with the constraint $n_s\mu_s+n_d\mu_d=s+d$. Taking the negative logarithm, losing constant terms, and introducing a Lagrange multiplier term for this constraint (with a new unknown ζ), this becomes:

$$-s\ln(\mu_s n_s)-d\ln(\mu_d n_d)+\frac{1}{2\sigma^2}\ln^2\frac{\mu_d}{\mu_s}+\zeta\cdot(n_s\mu_s+n_d\mu_d-s-d)\equiv F.$$

The function $F(\mu_s,\mu_d)$ can be easily minimized against μ_s and μ_d, yielding estimates of these quantities. To demonstrate this, first the derivatives of F are calculated and set to 0:

$$\mu_s\frac{\partial F}{\partial\mu_s}=\zeta\mu_s n_s-s-\frac{1}{\sigma^2}\ln\frac{\mu_d}{\mu_s}=0,\quad \mu_d\frac{\partial F}{\partial\mu_d}=\zeta\mu_d n_d-d+\frac{1}{\sigma^2}\ln\frac{\mu_d}{\mu_s}=0.$$

The sum of these, $\zeta\cdot(\mu_s n_s+\mu_d n_d)=s+d$, results in the value for $\zeta=1$. The derivative of F with respect to ζ gives back the constraint $n_s\mu_s+n_d\mu_d=s+d$, which yields an expression of μ_d as a function of μ_s. Plugging it into the first of the above two equations one gets

$$f=\mu_s\frac{\partial F}{\partial\mu_s}(\mu_s,\mu_d(\mu_s))=\mu_s n_s-s-\frac{1}{\sigma^2}\ln\frac{\mu_d(\mu_s)}{\mu_s}=0.$$

This equation can be solved with a few iterations of the Newton's root finding method starting with a solution to

$$\mu_s=\mu_d(\mu_s):\qquad \mu_s=\mu_d=\frac{s+d}{n_s+n_d}.$$

At each iteration the value of μ_s is adjusted by $-f/f'$, where the derivative is evaluated as

$$f'=n_s\left(1+\frac{1}{\sigma^2}(\frac{1}{\mu_s n_s}+\frac{1}{\mu_d n_d})\right).$$

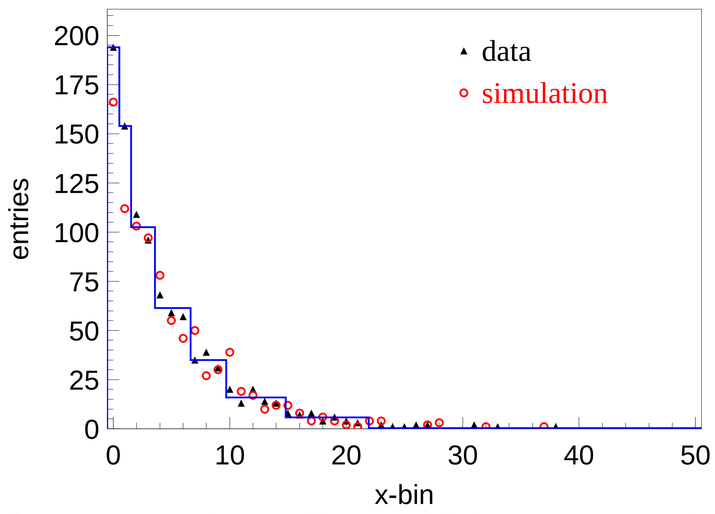
Once the likelihood function is solved for the best values of μ_s and μ_d, these can be plugged into the likelihood ratio given above. One can now write the likelihood ratio as a sum over all bins:

$$\sum_i\left[s_i\ln\frac{s_i/n_s}{\mu_s^i}+d_i\ln\frac{d_i/n_d}{\mu_d^i}+\frac{1}{2\sigma^2}\ln^2\frac{\mu_d^i}{\mu_s}\right].$$

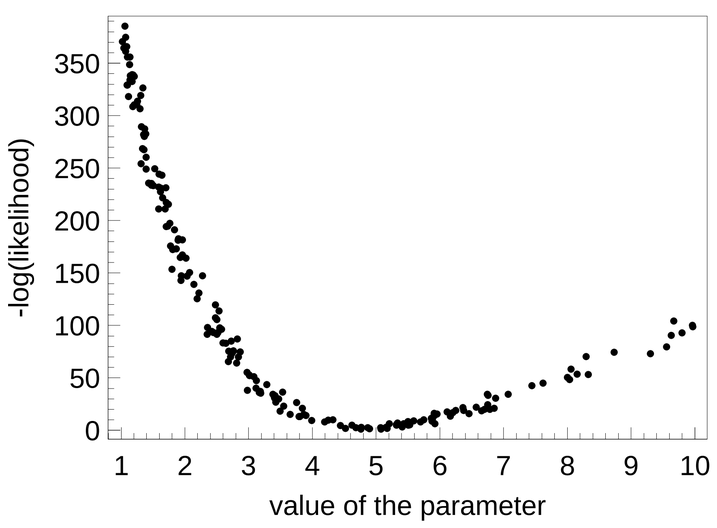
The probability $P(\text{same process})$ can also be thought of as a convolution of the binomial probability part of the expression with the penalty term. Solving the convolution integral approximately with the Laplace's method results (up to a constant term) in an expression for $P(\text{same process})$ given above.

Example

Assume the distribution of counts following the following PDF: $\exp(-x/\mu)/\mu$ with $\mu_0=5$.



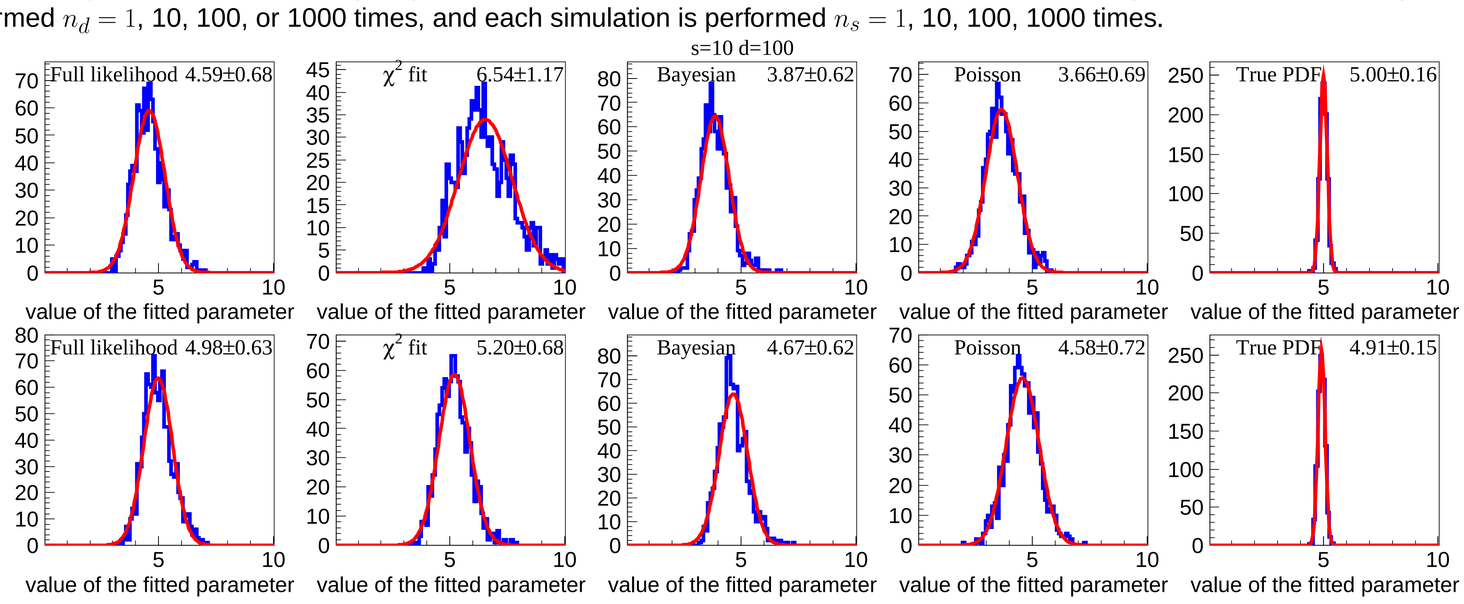
Data and simulation for $n_d=100$ and $n_s=100$. Also shown in blue solid line is the optimized binning applied as described in [1].



Likelihood values for random simulations in the vicinity of the true value of $\mu_0=5$.

Comparison with other methods

In this section we will compare the performance of parameter reconstruction using the method of this paper and three alternative methods: Poisson, χ^2, and Bayesian. We also reconstruct using the "true" Poisson likelihood by calculating the expectation in each bin directly from the probability distribution function (PDF). The sets are drawn with the mean total number of counts per set of 10. This "experiment" is performed $n_d=1$, 10, 100, or 1000 times, and each simulation is performed $n_s=1$, 10, 100, 1000 times.



Distributions of the reconstructed value of μ for 1000 random drawings of data with $n_d=100$. For each data set, 200 simulation sets with μ sampled log-uniformly in the interval $(\mu_0/5,\ 2\mu_0)$ are drawn with $n_s=10$ and the one with the best likelihood value is chosen to represent the result of the fit. The top row uses the bins of size 1, the bottom row uses the bins of optimized size. The Poisson "true PDF" likelihood is always the best, as expected, and does not depend on the binning method. In all other cases the result is improved with the optimized binning. The "full likelihood" corresponds to the likelihood description of this paper and gives the best result in this example.

The new likelihood gives the best result (closest to $\mu_0=5$ with smallest spread) more often than the other methods, except, of course, the method using the Poisson likelihood with the mean estimated from the analytical expression for the PDF. We also comment that the optimized binning improves the result in most cases.

		default binning					optimized binning				
n_s	n_d	Full	χ^2	Bayesian	Poisson	True PDF	Full	χ^2	Bayesian	Poisson	True PDF
1	1	3.79 ± 1.97	3.82 ± 1.98	$\mathbf{5.71\pm2.80}$	3.06 ± 2.30	5.02 ± 1.71	3.57 ± 2.22	3.57 ± 2.22	3.73 ± 2.43	$\mathbf{4.05\pm2.70}$	3.81 ± 2.44
1	10	3.53 ± 1.52	$\mathbf{5.15\pm1.80}$	$\mathbf{5.61\pm1.90}$	2.82 ± 1.98	5.00 ± 0.51	$\mathbf{5.05\pm1.80}$	5.29 ± 1.72	3.99 ± 1.75	2.61 ± 1.89	4.53 ± 0.49
1	100	3.46 ± 1.45	$\mathbf{5.66\pm1.88}$	2.35 ± 1.44	2.50 ± 1.64	5.00 ± 0.16	$\mathbf{4.95\pm1.69}$	7.00 ± 1.66	3.47 ± 1.65	2.76 ± 1.81	4.92 ± 0.15
1	1000	3.39 ± 1.43	$\mathbf{5.61\pm1.90}$	2.31 ± 1.33	2.47 ± 1.57	5.00 ± 0.07	3.62 ± 1.49	$\mathbf{5.86\pm1.92}$	2.30 ± 1.31	2.51 ± 1.61	4.99 ± 0.07
10	1	$\mathbf{4.48\pm1.80}$	2.06 ± 1.37	6.16 ± 2.97	1.48 ± 0.59	5.09 ± 1.72	3.42 ± 2.13	$\mathbf{3.48\pm2.20}$	3.46 ± 2.11	3.31 ± 2.18	3.71 ± 2.34
10	10	4.62 ± 0.94	$\mathbf{4.63\pm0.96}$	3.99 ± 1.26	2.40 ± 0.88	5.00 ± 0.50	$\mathbf{4.62\pm0.78}$	4.62 ± 0.78	4.43 ± 0.80	3.22 ± 1.48	4.53 ± 0.48
10	100	$\mathbf{4.59\pm0.67}$	6.54 ± 1.17	3.87 ± 0.62	3.66 ± 0.69	5.00 ± 0.16	$\mathbf{4.98\pm0.63}$	5.20 ± 0.68	4.66 ± 0.62	4.58 ± 0.72	4.91 ± 0.15
10	1000	$\mathbf{4.64\pm0.65}$	8.61 ± 0.98	3.85 ± 0.57	3.96 ± 0.65	5.00 ± 0.06	$\mathbf{5.01\pm0.71}$	7.05 ± 1.14	4.26 ± 0.62	4.28 ± 0.75	4.99 ± 0.06
100	1	$\mathbf{4.78\pm1.75}$	1.52 ± 1.14	6.65 ± 3.11	2.09 ± 1.21	5.00 ± 1.69	3.48 ± 2.16	$\mathbf{3.63\pm2.35}$	3.53 ± 2.22	3.47 ± 2.17	3.90 ± 2.52
100	10	$\mathbf{4.86\pm0.65}$	3.35 ± 0.51	4.41 ± 0.92	4.05 ± 0.89	5.01 ± 0.52	$\mathbf{4.55\pm0.56}$	4.48 ± 0.63	4.44 ± 0.57	4.34 ± 0.98	4.54 ± 0.50
100	100	4.94 ± 0.35	$\mathbf{4.95\pm0.34}$	4.63 ± 0.34	4.72 ± 0.36	5.00 ± 0.16	$\mathbf{4.92\pm0.27}$	4.92 ± 0.27	4.85 ± 0.27	4.92 ± 0.26	4.92 ± 0.16
100	1000	$\mathbf{4.93\pm0.26}$	5.39 ± 0.33	4.71 ± 0.26	4.86 ± 0.28	5.00 ± 0.06	4.99 ± 0.24	5.09 ± 0.25	4.90 ± 0.24	$\mathbf{5.00\pm0.25}$	4.99 ± 0.06
1000	1	$\mathbf{4.87\pm1.66}$	1.44 ± 1.09	6.88 ± 3.21	3.11 ± 1.77	4.94 ± 1.67	3.33 ± 2.12	$\mathbf{3.53\pm2.39}$	3.51 ± 2.32	3.33 ± 2.12	3.82 ± 2.46
1000	10	$\mathbf{4.94\pm0.56}$	2.25 ± 0.37	4.51 ± 0.82	4.72 ± 0.70	5.01 ± 0.50	$\mathbf{4.52\pm0.51}$	4.38 ± 0.75	4.42 ± 0.53	4.49 ± 0.60	4.53 ± 0.49
1000	100	$\mathbf{4.98\pm0.23}$	4.46 ± 0.23	4.76 ± 0.23	4.93 ± 0.24	4.99 ± 0.17	$\mathbf{4.91\pm0.20}$	4.86 ± 0.21	4.87 ± 0.19	4.91 ± 0.20	4.91 ± 0.16
1000	1000	$\mathbf{4.99\pm0.13}$	4.99 ± 0.13	4.92 ± 0.13	4.98 ± 0.13	5.00 ± 0.06	4.99 ± 0.11	4.99 ± 0.11	4.97 ± 0.11	$\mathbf{5.00\pm0.11}$	4.99 ± 0.06

Values of μ reconstructed by various methods. Intervals are shown with mean and rms of the reconstructed values for 1000 of random data drawings. The result of the first four methods (i.e., except Poisson with true PDF shown in the last column) closest to $\mu_0=5$ is shown with bold font. For all five methods intervals that do not enclose the true value of $\mu_0=5$ (i.e., those that are significantly biased) are shown with italic font. The column designated with "Full" corresponds to the likelihood description of this paper.

References

[1] J. Scargle, Bayesian blocks algorithm: http://trotsky.arc.nasa.gov/~jeffrey/, 2012.