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 ratio is the "likelihood" that should be used for comparison of different simulation sets with data. The denominator in the above expression effectively factors out the dependence of the probability P on the number of "states" (i.e., possible values of s) around the most likely value of s. (This dependence comes in because of the normalization condition: the sum of P over all states should be 1, which means that the value of P for the most likely values of s is lower for larger $s+d$.)

Binning

If the data is binned in equal-length consecutive bins one may occasionally benefit from combining the counts in contiguous subsets of bins into larger bins of length L (L is the number of initial-size small bins in the larger bin). The benefit of combining the counts of several bins must be weighted against the possibility of measuring the change in rate between the bins. If the counts are comparable to each other or are very small in the two adjacent bins (i.e., are within the statistical uncertainties), then the rate change across the these bins is consistent with zero, and thus the bins should be combined. This is the principal idea behind the "Bayesian blocks" algorithm of [1]. Note how the probability ratio construction used by this algorithm is quite similar to the probability ratio used in this report.
The multinomial probability for a particular distribution of counts in bins under consideration is

$$P(\{\mu_i\};\{d_i\}|\sum_i d_i=D)=\frac{D!}{\prod_i d_i!}\cdot\mu_i^{d_i}.$$

The first hypothesis is that the counts in these bin are independent from each other. It follows that $\mu_i=d_i$ for all i. The second hypothesis is that the counts in all of the bins under consideration are governed by the process with the same rate $\mu=\mu_i$ for all i. If follows that $\mu=D/L$, where L is the number of bins.
To determine whether the bins should be combined or be considered separately we form the likelihood ratio:

$$\frac{P(\text{one bin})}{P(\text{many bins})}=\frac{(D/L)^D}{\prod_i d_i^{d_i}},\quad\text{where}\quad D=\sum_i d_i.$$

Now we note that this ratio never exceeds 1, as the value of the likelihood maximized with a constraint $\mu=\mu_i$ is necessarily smaller or equal to than the value of the likelihood maximized without any restrictions.
Note that for any permutation of bins i the ratio above does not change. However, with the assumption of the same rate across the bins all ($L!$) of the permutations are equivalent to merely being different statistical realizations of the same experiment, all leading to the same term $\prod_i d_i^{d_i}$ in the denominator. On the other hand there is only one possible realization of the data arrangement that leads to the term $(D/L)^D$ in the nominator.
We will now make an intuitive conclusion that it is fair to decide whether to combine the bins i into a single large bin with length L based on a comparison of the ratio above with $1/L!$. We can easily generalize this rule for comparing various arrangements of bins k of different sizes L_k: we choose the arrangement that leads to the largest sum (taking the log of the probability ratio)

$$\sum_k D_k\cdot\ln\left(\frac{D_k}{L_k}\right)+\ln(L_k!).$$

To get a little more insight into the expression given above we write it for a number L of bins i each with approximately the same counts $d_i=d+\delta_i$, $D=\sum_i d_i=d\cdot L$, $\delta_i/d\ll1$:

$$\sum_i d_i\cdot\ln d_i+\ln(1!)=\sum_i(d+\delta_i)\ln(d+\delta_i)\approx\sum_i\left(d\ln d+\frac{\delta_i^2}{2d}\right)=D\ln\left(\frac{D}{L}\right)+\frac{1}{2}\sum_i\frac{\delta_i^2}{d}.$$

We compare this with the expression for a single large bin of length L:

$$D\ln\left(\frac{D}{L}\right)+\ln(L!).$$

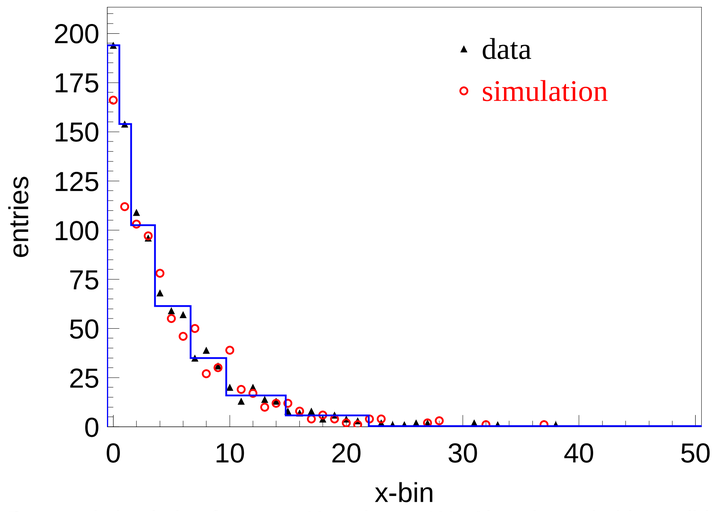
So, the decision to combine the bins is based on whether

$$\frac{1}{2}\sum_i\frac{\delta_i^2}{d}<\ln(L!)\approx L\cdot\ln\left(\frac{L}{e}\right)\sim L,\quad\text{or}\quad\sum_i\frac{\delta_i^2}{d}\lesssim L,$$

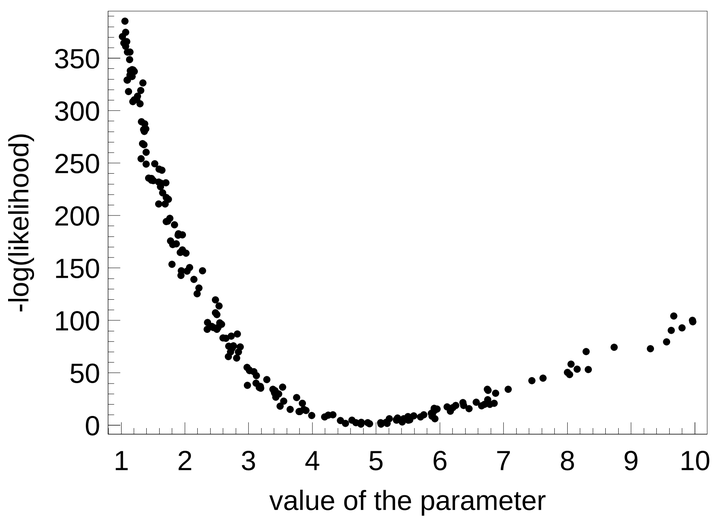
i.e. when the χ^2 (i.e., goodness-of-fit) of description of data with the mean d in the bins is on the order of or smaller than the number of bins (i.e., degrees of freedom of the χ^2).

Example

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



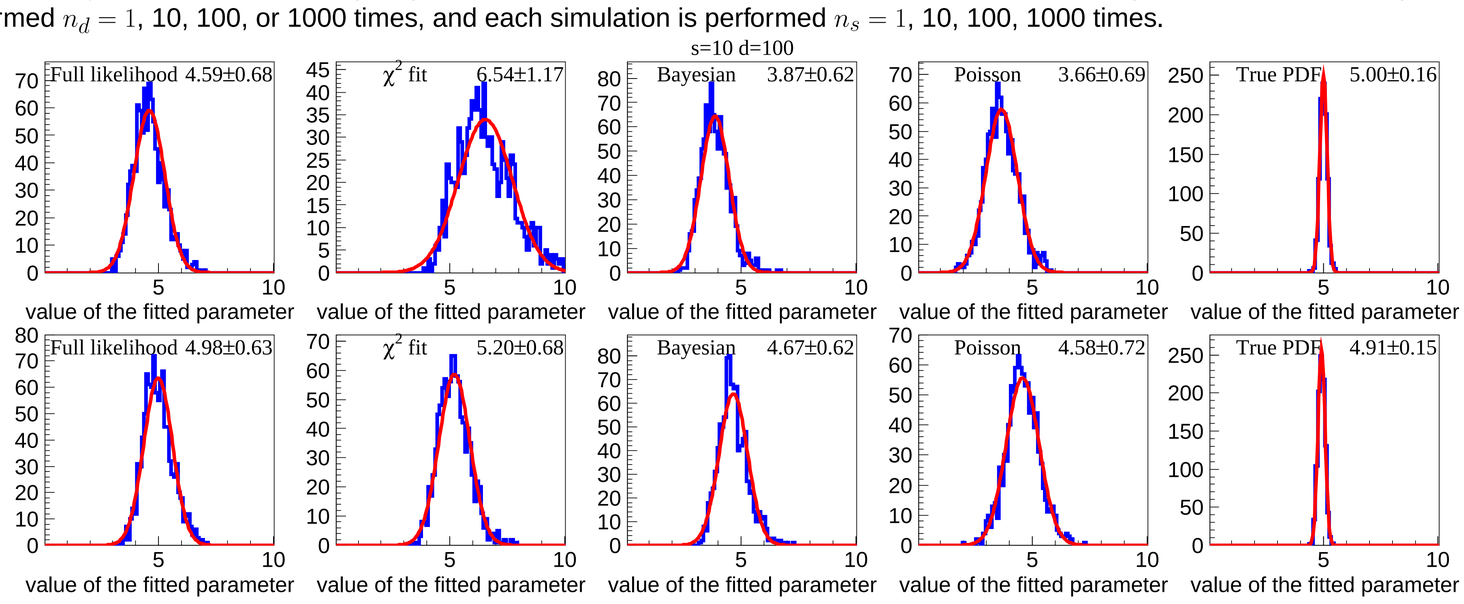
Data and simulation for $n_d=100$ and $n_s=100$. Also shown in blue solid line is the optimized binning.



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	$\mathbf{5.15\pm1.80}$	$\mathbf{5.28\pm1.98}$	$\mathbf{5.28\pm1.88}$	5.00 ± 0.51	$\mathbf{5.05\pm1.80}$	5.29 ± 1.72	3.99 ± 1.75	$\mathit{2.61\pm1.89}$	4.53 ± 0.49
1	100	$\mathit{3.46\pm1.45}$	$\mathbf{5.66\pm1.88}$	$\mathit{2.35\pm1.44}$	$\mathit{2.50\pm1.64}$	5.00 ± 0.16	$\mathbf{4.95\pm1.69}$	$\mathit{7.00\pm1.66}$	3.47 ± 1.65	$\mathit{2.76\pm1.81}$	4.92 ± 0.15
1	1000	$\mathit{3.39\pm1.43}$	$\mathbf{5.61\pm1.90}$	$\mathit{2.31\pm1.33}$	$\mathit{2.47\pm1.57}$	5.00 ± 0.07	3.62 ± 1.49	$\mathbf{5.86\pm1.92}$	$\mathit{2.30\pm1.31}$	$\mathit{2.51\pm1.61}$	4.99 ± 0.07
10	1	$\mathbf{4.48\pm1.80}$	$\mathit{2.06\pm1.37}$	6.16 ± 2.97	$\mathit{1.48\pm0.59}$	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	$\mathit{2.40\pm0.88}$	5.00 ± 0.50	$\mathbf{4.62\pm0.78}$	4.62 ± 0.78	4.43 ± 0.80	$\mathit{3.22\pm1.48}$	4.53 ± 0.48
10	100	$\mathbf{4.59\pm0.67}$	$\mathit{6.54\pm1.17}$	$\mathit{3.87\pm0.62}$	$\mathit{3.66\pm0.69}$	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	$\mathbf{4.64\pm0.65}$	$\mathit{8.61\pm0.98}$	$\mathit{3.85\pm0.57}$	$\mathit{3.96\pm0.65}$	5.00 ± 0.06	$\mathbf{5.01\pm0.71}$	$\mathit{7.05\pm1.14}$	$\mathit{4.26\pm0.62}$	4.28 ± 0.75	4.99 ± 0.06
100	1	$\mathbf{4.78\pm1.75}$	$\mathit{1.52\pm1.14}$	6.65 ± 3.11	$\mathit{2.09\pm1.21}$	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	$\mathbf{4.86\pm0.65}$	$\mathit{3.35\pm0.51}$	4.41 ± 0.92	$\mathit{4.05\pm0.89}$	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	$\mathbf{4.95\pm0.34}$	$\mathit{4.63\pm0.34}$	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	$\mathbf{4.93\pm0.26}$	$\mathit{5.39\pm0.33}$	$\mathit{4.71\pm0.26}$	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	$\mathbf{4.87\pm1.66}$	$\mathit{1.44\pm1.09}$	6.88 ± 3.21	$\mathit{3.11\pm1.77}$	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	$\mathbf{4.94\pm0.56}$	$\mathit{2.25\pm0.37}$	4.51 ± 0.82	4.72 ± 0.70	5.01 ± 0.50	$\mathbf{4.52\pm0.51}$	4.38 ± 0.75	$\mathit{4.42\pm0.53}$	4.49 ± 0.60	4.53 ± 0.49
1000	100	$\mathbf{4.98\pm0.23}$	$\mathit{4.46\pm0.23}$	$\mathit{4.76\pm0.23}$	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.