Jianming Qian
January 13, 2001
DØ note 3823

Comments on Unbinned and Binned Likelihood Functions

The purpose of this short note is to point out a subtle difference between unbinned and binned (based on Poisson probability) likelihood functions. I came to realize this issue while reviewing the W decay width (Γ_W) analysis. In this analysis, the transverse mass (M_T) spectrum is fitted in a range different from the range in which the M_T spectrum is normalized. This difference introduces an unwanted Γ_W-dependent term in the binned likelihood function, as explained below.

Let $f(x;a)$ be the probability density function of a directly measurable quantity x, where a is the parameter to be extracted using the maximum likelihood method. In most cases, $f(x;a)$ is determined using Monte Carlo and is normalized to unity within a range of x, say between $[x_L, x_H]$ (normalization range):

$$\int_{x_L}^{x_H} f(x;a)dx = 1$$

However, it is often the practice to carry out the fit over a more restrict range of x, either to maximize discrimination for different values of a or to minimize systematic error. Let's assume that there are a total of N_0 events with $x_L < x < x_H$ and N of them have the measured x values $\{x_1, x_2, ..., x_{N-1}, x_N\}$ within the fitting range $[x_l, x_h]$ ($x_L < x_l,\ x_h < x_H$). The unbinned likelihood function (according to Particle Data Group) is:

$$\mathcal{L}_u = \prod_i f(x_i;a) \quad \text{or} \quad \log\mathcal{L}_u = \sum_i \log f(x_i;a)$$

where the product $\prod_i$ and the summation $\sum_i$ run over the N events in the fitting range. For simplicity, I have ignored backgrounds in the likelihood. The parameter a can then be extracted by maximizing $\mathcal{L}_u$ (or $\log\mathcal{L}_u$).

In practice, one often bins the measurement first, particularly when the number of events is large. Let (y_j, δ_j, n_j) be the central value, the width, and the entry of bin j, the number of events expected in the bin is given by

$$\mu_j = p_j N_0 = [f(y_j;a)\delta_j]N_0$$

where $p_j = f(y_j;a)\delta_j$ is the probability to have x in bin j. The Poisson-based binned likelihood is

$$\mathcal{L}_b = \prod_j \frac{e^{-\mu_j}}{n_j!}{\mu_j}^{n_j}$$

$$\begin{aligned}\log\mathcal{L}_b &= \textstyle\sum_j [n_j \log\mu_j - \mu_j - \log(n_j!)] \\ &= \textstyle\sum_j \{n_j \log[f(y_j;a)\delta_j N_0] - f(y_j;a)\delta_j N_0 - \log(n_j!)\} \\ &= \textstyle\sum_j \{n_j \log f(y_j;a) - f(y_j;a)\delta_j N_0 + n_j\log(\delta_j N_0) - \log(n_j!)\}\end{aligned}$$

where j runs over all bins in the fitting range. For the purpose of extracting a, the last two terms (independent of a) of the above logarithmic likelihood can be dropped:

$$\log\mathcal{L}_b = \sum_j n_j \log f(y_j;a) - N_0 \sum_j f(y_j;a)\delta_j$$

Comparing the unbinned and binned likelihoods, we note that

$$\sum_i \log f(x_i;a) \approx \sum_j n_j \log f(y_j;a)$$

if the bin widths are reasonably small. Therefore

$$\log\mathcal{L}_b \approx \log\mathcal{L}_u - N_0 \sum_j f(y_j;a)\delta_j$$

The two likelihoods differ by a term dependent on the parameter to be extracted. Consequently the unbinned and binned approach will generally lead to different results.

This problem is caused by fitting over a more restrict range than the normalization range. It goes away if the two ranges are the same. In this case,

$$\sum_j f(y_j;a)\delta_j = 1 \;\; \{\int_{x_L}^{x_H} f(x;a)dx = 1\}$$

The unbinned and binned fits should therefore give identical results barring possible binning effects.

Alternatively one could define the binned likelihood to be

$$\mathcal{L}'_b = \prod_j p_j^{n_j} \Rightarrow \log \mathcal{L}'_b = \sum_j n_j \log p_j = \sum_j n_j \log f(y_j;a) + \sum_j n_j \log(\delta_j)$$

instead of using Poisson probability. By dropping the $a-$independent term, one gets

$$\log \mathcal{L}'_b = \sum_j n_j \log f(y_j;a)$$

which is binned calculation of the unbinned likelihood.