Uncertainties in Weighted Monte Carlo Data

R. Ellsworth

27 May 03

1 Introduction

The Corsika-Geant Milagro Monte Carlo program, developed by Stephan and Julie throws events on the detector with a distribution which is flat in radius. The distribution of actual shower axis hits (not necessarily triggers) at detector level is one of uniform density (ρ), so the distribution in r is

$$\frac{dn}{dr} = 2\pi r\rho$$

The reason for throwing flat in r, rather than linearly, is to increase the fraction of Monte-Carlo generated events which produce triggers. To compensate for the biasing, events are put in weighted histograms; each event is given a weight r.

This works fine for computing averages, such as trigger rates.

But other uses of the Monte Carlo data can require knowledge of the actual statistical error of a subset of events. So an important question is: given a Monte-Carlo-produced histogram of some quantity, what are the correct statistical errors for each bin of the histogram?

2 Uncertainties

In the jth bin of a histogram of some arbitrary quantity, the weighted histogram has, say, N_j events. Because each event is weighted by r,

$$N_j = \sum_{k=1}^{n_j} r_k$$

in which n_j is the actual number of **unweighted** Monte-Carlo events in the jth bin, and r_k is the core radius of the kth event in the jth bin. This is actually what is done; but the above equation is can be written in a way which is dimensionally correct:

$$N_j = \frac{1}{r_0} \sum_{k=1}^{n_j} r_k$$

in which $r_0 = 1\ cm$, and r is in cm.

We may safely assume that the fluctuations in n_j are Poisson. For the jth bin, it is also meaningful to consider the average radius of the events in it. This is just

$$\bar{r_j} = \frac{\sum_{k=1}^{n_j} r_k}{n_j} = r_0 N_j / n_j \tag{1}$$

In other words,

$$N_j = \frac{1}{r_0}\bar{r}_j n_j$$

From error propagation, the statistical uncertainty in N_j is then

$$\sigma_N^2 = \frac{1}{r_0^2}(\bar{r}^2\sigma_n^2 + n^2\sigma_{\bar{r}}^2)$$

Now $\sigma_n^2 = n$, and, using the central limit theorem

$$\sigma_{\bar{r}}^2 = \sigma_r^2/\sqrt{n}$$

in which σ_r^2 refers to the radial distribution of the events in the jth bin. So the result becomes

$$\sigma_N^2 = \frac{n}{r_0^2}(\bar{r}^2 + \sigma_r^2) \tag{2}$$

Note that $\bar{r}$ and σ_r are from unweighted data. This can be simplified using $\sigma_r^2 = \overline{r^2} - \bar{r}^2$ to

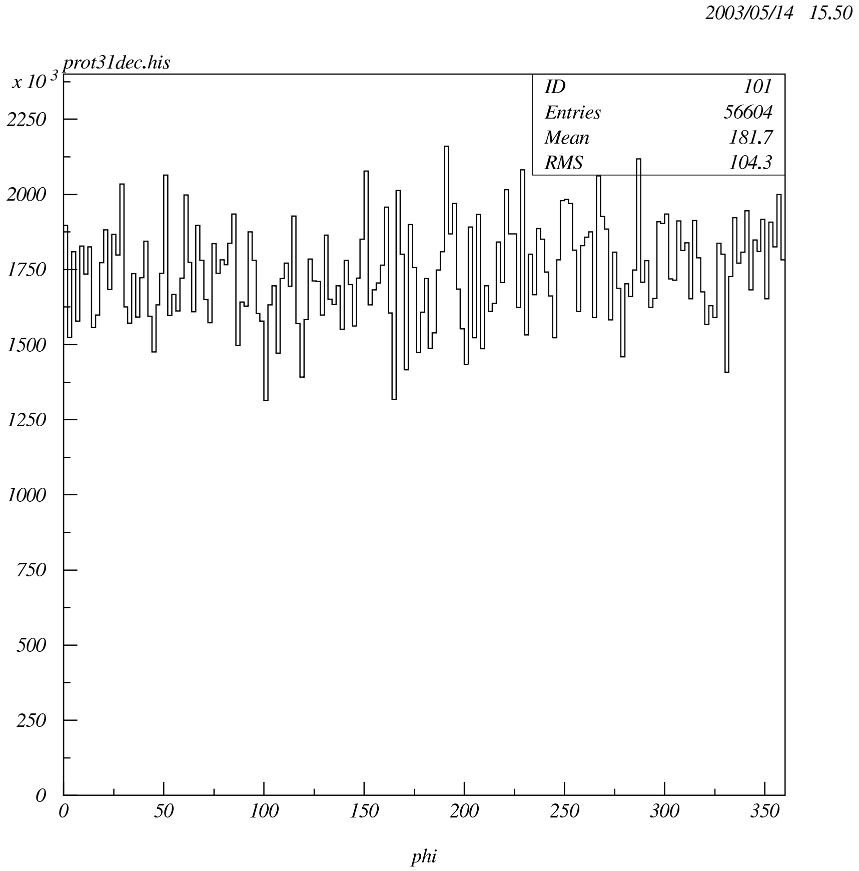
$$\sigma_N^2 = \frac{n\overline{r^2}}{r_0^2} \tag{3}$$

A useful approximation is possible when the quantity being binned is uncorrelated or weakly correlated with the core radius. In that case, we may obtain $\bar{r}$ and σ_r directly from the radial distribution of triggered, unweighted events.

This equation for the uncertainty should also permit the combination of weighted and unweighted event sets.

3 Test with Monte Carlo data

To test the above result, we need some binned quantity which is sampled many times, so we can study its fluctuations. The current Monte Carlo provides this in the azimuth distribution of weighted, triggered events. This is shown in the figure below.



There were $56,604$ triggered events in the Monte Carlo sample. The azimuth is the reconstructed one. Assuming a uniform ϕ distribution, the number in each of the 180 bins in this histogram is a sample of the same quantity. [1]

The average, over 180 bins, for number of weighted events per bin is

$$\bar{N} = 1738864$$

The rms deviation of this is found to be

$$\sigma(N) = 161,064 \tag{4}$$

(This was obtained by extracting the histogram to a data file, and computing the rms deviation from the mean of the number in each bin. The statistics given by PAW are in ϕ, not $N(\phi)$.) This $\sigma(N)$ is much larger than $\sqrt{\bar{N}}$, which is 1319. It is also larger than $\bar{r}\sqrt{\bar{n}}$, which is $110,655$.

Now we try to predict the result of Equation 4, using Equation 2. The average number of unweighted events per bin is $n = 314.5$ Using Equation (1) , the ensemble average of $\bar{r}$ is $\overline{\bar{r}} = 5529.6$ From the unweighted distribution of the radius for triggered events, the rms deviation of r is

$$\sigma(r) = 7,475$$

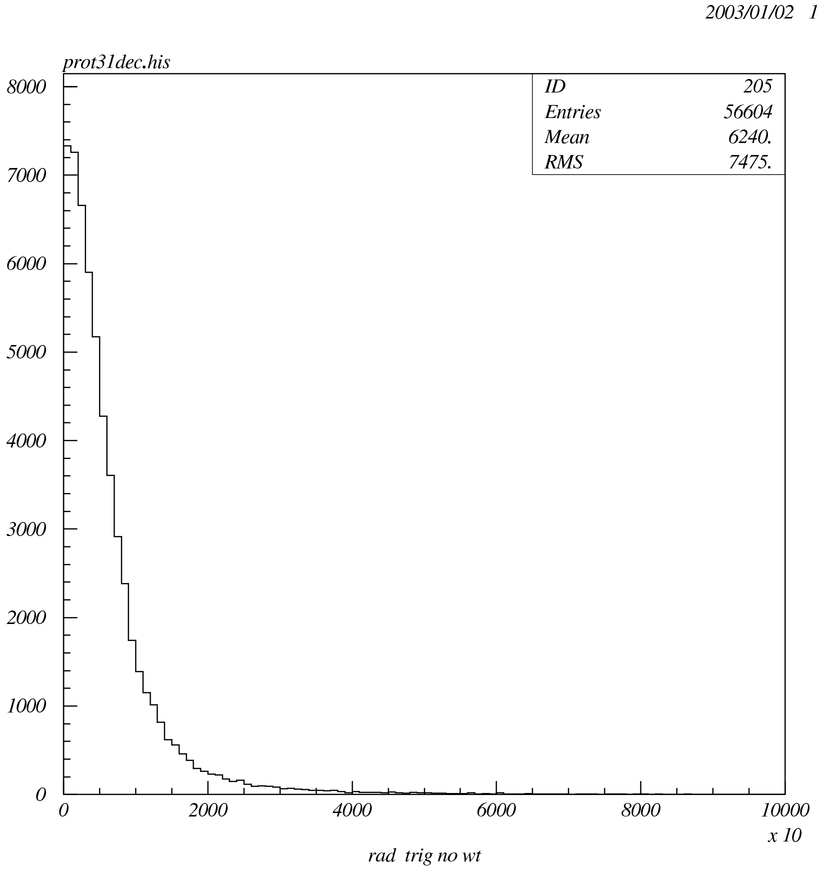
[1] The small assymetry in the ϕ distribution is being neglected here.

Combining these with equation 2 gives a predicted $\sigma(N)$

$$\sigma(N) = 164,883$$

which is within about 2% of the observed value.

Actually, this result is too good. Here is the radial distribution of triggered events, unweighted:



$\sigma(r)$ is indeed 7,475, but the average is not 5,530 but 6,240. Using the latter instead of 5,530 gives a result

$$\sigma(N) = 172,672$$

which is about 7% high.

I am not yet sure of the origin of this difference.