Error Estimates for the Bin Contents of a Convolution Histogram

Luc Demortier[1] and Luca Lista[2]

[1] *Laboratory of Experimental High-Energy Physics, Rockefeller University*

[2] *INFN, Napoli*

February 12, 2010

Abstract

We examine a generalization of a Monte Carlo procedure commonly used to model the missing transverse energy spectrum of $W \to \tau\nu$ events. This procedure introduces correlations between spectrum bins, thereby invalidating the usual multinomial error structure. We compute the errors on the bin contents from first principles.

1 Problem Setup

A common technique for generating the distribution of a random variable Z that is itself a function $c(\vec{X}, \vec{Y})$ of two possibly multidimensional, possibly dependent random variables $\vec{X}$ and $\vec{Y}$, is the following:

1. Draw a number n of variates distributed as $\vec{X}$;
2. For each $\vec{X}$:
   (a) Draw a number k of variates distributed as $\vec{Y}$ given $\vec{X}$;
   (b) Compute $Z = c(\vec{X}, \vec{Y})$;
3. Histogram the Z's.

This technique can be applied to the modeling of the missing transverse energy spectrum of events containing $W \to \tau\nu$ decays, where one must combine contributions from the W decay neutrino and from the tau lepton. A typical procedure is to substitute Monte Carlo tau decays for the muons in a data sample of $W \to \mu\nu$ events. When the size of the muon data sample is limited, the precision of the result can be increased by substituting several tau decays for each available muon. However, this introduces nontrivial correlations between spectrum bins, because a given $W \to \mu\nu$ event can now contribute to more than one bin (only in the special case $k = 1$ are the correlations purely multinomial).

The question we wish to address is how to calculate the errors on the bin contents of the Z histogram in general. We derive these errors in the next section.

2 Solution

We adopt the standard statistical convention of designating random variables with capital letters and their observed values with the corresponding small-case letters. Let $f(\vec{x})$ be the probability density of $\vec{X}$ and $g(\vec{y}\,|\,\vec{x})$ that of $\vec{Y}$ given $\vec{X}$. Define further:

$$\begin{aligned} M_{ab} &\equiv \text{number of } Z \text{ values in bin } [z_a, z_b] \text{ coming from a given } \vec{X} \\ &\quad \text{(observed values of } M_{ab} \text{ will be written as } m_{ab}(\vec{x}) \text{ to indicate} \\ &\quad \text{their association with observed values of } \vec{X}); \\ N_{ab} &\equiv \text{total number of } Z \text{ values in } [z_a, z_b]: \quad n_{ab} \;=\; \sum_{i=1}^{n} m_{ab}(\vec{x}_i); \qquad (2.1) \\ n'_{ab} &\equiv \sum_{i=1}^{n} [m_{ab}(\vec{x}_i)]^2. \end{aligned}$$

The crucial insight is that within a given Z bin $[z_a, z_b]$ the M_{ab} are independent, since each of them is obtained by drawing a new $\vec{X}$ from $f(\vec{x})$ and k new $\vec{Y}$'s from $g(\vec{y}\,|\,\vec{x})$. In addition, the M_{ab} are identically distributed since the same f and g are used throughout. We therefore have, from Eq. (2.1):

$$\mathrm{Var}(N_{ab}) \;=\; n\,\mathrm{Var}(M_{ab}). \qquad (2.2)$$

An unbiased estimate of the expectation $\mathbb{E}(M_{ab})$ is n_{ab}/n, and the corresponding unbiased estimate of $\mathrm{Var}(M_{ab})$ is then:

$$\mathrm{Var}(M_{ab}) \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\left(m_{ab}(\vec{x}_i) \;-\; \frac{n_{ab}}{n}\right)^2 \;=\; \frac{1}{n-1}\left(n'_{ab} \;-\; \frac{n_{ab}^2}{n}\right). \qquad (2.3)$$

Therefore:

$$\mathrm{Var}(N_{ab}) \;=\; \frac{n}{n-1}\left(n'_{ab} \;-\; \frac{n_{ab}^2}{n}\right). \qquad (2.4)$$

Note that if $k = 1$ all the $m_{ab}(\vec{x}_i)$ are either 0 or 1, so that $n'_{ab} = n_{ab}$ and the above formula reduces to the standard binomial case. More importantly, Eq. (2.4) expresses $\mathrm{Var}(N_{ab})$ in terms of quantities that are directly accessible when making the histogram; there is no need to run large ensembles of toy experiments to figure out the uncertainties on the bin contents. The histogramming procedure could be something like this:

```
1   Initialize histograms H1, H2, and H3 with B bins from z_min to z_max.
2   For i = 1, ..., n:
3       Generate x_i ~ f(x).
4       Zero histogram H3.
5       For j = 1, ..., k:
6           Generate y_ij ~ g(y | x_i).
7           Set z_ij = c(x_i, y_ij).
8           Histogram z_ij in H1.
9           Histogram z_ij in H3.
10      Add the square of H3 to H2, bin by bin.
11  Compute H4 ≡ sqrt( n/(n-1) [H2 - H1^2/n] ) (also bin by bin).
```

Histogram H_4 then contains the estimated errors on the bins of histogram H_1.

3 Example

We illustrate our result with an example that is somewhat inspired by the $W \to \tau\nu$ problem. Suppose $\vec{X}$ and $\vec{Y}$ are two-dimensional vectors generated as follows:

$$\vec{X} \equiv (X_1, X_2) \equiv (U\cos S,\, U\sin S): \qquad U \sim \frac{1}{\tau}e^{-u/\tau}, \qquad S \sim \mathcal{U}[0, 2\pi[, \tag{3.1}$$

$$\vec{Y} \equiv (Y_1, Y_2) \equiv (V\cos T,\, V\sin T): \qquad V\,|\,U \sim \mathcal{N}(u, \sigma^2), \quad T \sim \mathcal{U}[0, 2\pi[, \tag{3.2}$$

where τ and σ are known constants. In words, $\vec{X}$ has an exponentially distributed magnitude, $\vec{Y}$ a Gaussian distributed magnitude conditional on the magnitude of $\vec{X}$, and both have uniform azimuth. One could think of $\vec{X}$ as the missing transverse energy of the neutrino from W decay and $\vec{Y}$ as that of the tau lepton, although the above distributions are motivated by convenience rather than by any resemblance with real physics processes. The combination rule $c(\vec{X}, \vec{Y})$ is the magnitude of $\vec{X} + \vec{Y}$:

$$Z = c(\vec{X}, \vec{Y}) = \sqrt{(X_1 + Y_1)^2 + (X_2 + Y_2)^2} = \sqrt{U^2 + V^2 + 2UV\cos(S - T)}. \tag{3.3}$$

We generated 50,000 toy experiments with $\tau = \sigma = 1$, $n = 10$, and $k = 100$. The results are displayed in Fig. 1. The left panel shows agreement between the toy experiment averages of the estimated variances of the bin contents and the toy experiment variances of the bin contents, indicating that the variance estimated with eq. (2.4) is unbiased. The right panel shows the same with standard deviation instead of variance. Here we do see some bias at low bin contents, where the estimated standard deviation underestimates the "true" standard deviation. This behavior is expected since bias is not invariant under nonlinear transformations such as the square root. Furthermore, the square root is a concave function, so that by Jensen's inequality the estimated standard deviation should indeed *underestimate*.

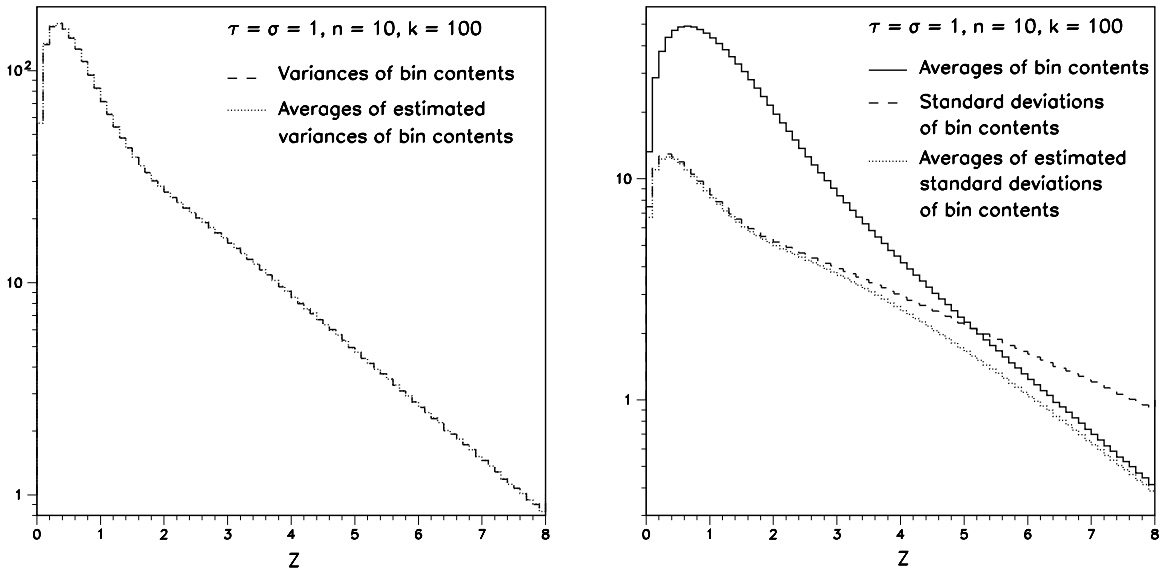

Figure 1: Result of 50,000 toy experiments with $\tau = \sigma = 1$, $n = 10$, and $k = 100$.