CDF/DOC/JET/PUBLIC/3126
Version 3.0

Correlations from Systematic Corrections to Poisson-Distributed Data in Log-Likelihood Functions

T. Devlin

August 19, 1995

1 Introduction

CDF data analyses are often confronted with systematic corrections to a set of data points, e.g. the jet inclusive cross section, which give rise to correlated uncertainties between points. With Gaussian statistics in χ^2 fits and tests of significance, these correlations can be included through the variance matrix.[1, 2] However, when the uncertainties obey Poisson statistics, a likelihood function is more appropriate, and nothing corresponding to the variance matrix exists. At times I have heard colleagues express despair about accounting properly for these correlations in tests of significance when comparing with theoretical predictions. Take heart, friends; there is a way to do it. I will outline it in Sec. 2

Extension of the same ideas to deal with backgrounds is discussed in Sec. 3.

Another area of concern is that of efficiency corrections. Usually, efficiencies are evaluated as the fraction of some set of events which satisfy some criterion such as a trigger or a cut. The result is restricted to the range (0.0 : 1.0), and it obeys binomial statistics.

Most people evaluate the efficiency and its uncertaintly correctly from the binomial statistics, but from then on it is treated as if it obeys Gaussian statistics and combined in quadrature with the uncertainty from counting statistics. When the efficiency is near 100%, and when its uncertainty dominates the overall uncertainty, this can be inaccurate. I will discuss a simple method for dealing with this properly in Sec 4. There are certain experiments in which the "signal" obeys binomial statistics, and this treatment is applicable to such situations.

2 Poisson Statistics and Correlations

To discuss Poisson statistics and correlated systematic uncertainties, let us define a data set and the statistical uncertainties. For our purposes this can include any uncertainties for which there is no correlation between data points. Assume that the data originate as a histogram, i.e. the number of events in each of a set of bins, denoted by subscript i,

- n_i = number of events in the i^{th} bin
- the r.m.s. deviation of n_i is $\Delta n_i = \sqrt{n_i + 1}$. Using this in a Gaussian "approximation" is usually reasonable, but even for large n_i it can result in significant biases in certain situations, e.g. ratios of such numbers.[4]

- $Y_i = B_i n_i$ = the physics quantitity in the i^{th} bin, e.g. the jet inclusive cross section, $d\sigma/dE_t$. B_i includes the luminosity, geometric acceptance (such as the cuts on pseudo-rapidity, $\Delta\eta$), the trigger efficiency, the bin width (ΔE_t), and any other factors needed to compute the cross section from the raw number of events.
- $\sigma_i = \Delta Y_i = B_i\sqrt{n_i}$ = the "Gaussian" uncertainty in Y_i.
- T_i = a theoretical prediction for Y_i.
- $t_i = T_i/B_i$ = the corresponding theoretical prediction for n_i.

In what follows, it is instructive to recall for Gaussian statistics the exact relationship between χ^2 and the likelihood function:

$$\chi^2 = -2\ln[\mathcal{L}/\mathcal{L}_0] \tag{1}$$

where, ignoring systematics,

$$\mathcal{L} = \prod_i \left(\frac{1}{\sigma_i\sqrt{2\pi}}\right) \exp^{-\frac{(Y_i-T_i)^2}{2(\sigma_i)^2}} \tag{2}$$

and

$$\mathcal{L}_0 = \prod_i \left(\frac{1}{\sigma_i\sqrt{2\pi}}\right). \tag{3}$$

Note that the sum of exponents in Eq. 2 is just $-\chi^2/2$. Eq. 3 corresponds to a perfect fit, i.e. $Y_i = T_i$ and $\chi^2 = 0$.

We assume that the B_i include the "best" estimate of a number of systematic corrections to the data such as overall normalization to luminosity, energy scale corrections, unsmearing, etc. For each of these, evaluate the fractional change for ± 1 std. dev. changes away from the best value. For any single type of systematic correction, represented by subscript k, this changes Y_i by a factor

$$(1 + f_i^k S_k)$$

This defines the f_i^k, and we introduce the parameter, S_k, where $S_k = 0$ for the "best" correction, and $S_k = \pm 1$ for the one-std.-dev limits on the correction. In a fit in which S_k are allowed to vary, the data points as free to vary by the multiplicative factors:

$$Y_i \rightarrow Y_i \prod_k (1 + f_i^k S_k) \tag{4}$$

Each S_k is subject to the constraint that the "best" estimate of its value is $S_k = 0.0$, with an uncertainty, $\Delta S_k = 1.0$.

For Gaussian uncertainties, we can incorporate the effects of uncertainties in these correlated corrections as follows:

$$\chi^2 = \sum_i \frac{(Y_i \prod_k (1 + f_i^k S_k) - T_i)^2}{(\Delta Y_i)^2} + \sum_k S_k^2$$

where, to dispell any confusion, we note that

$$\sum_k S_k^2 = \sum_k \frac{(S_k - S_{0k})^2}{(\Delta S_k)^2} \tag{5}$$
$$S_{0k} = 0.0$$
$$\Delta S_k = 1.0$$

In the variance-matrix method,[1, 2] the f_i^k are used to compute appropriate contributions to the variance matrix for fitting, and, statistically, that method is entirely equivalent to the one described here.

If the data are Poisson-distributed, there is no place for a variance matrix. We can define a likelihood function as in Ref. 1, $\mathcal{L}$, and the equivalent perfect fit, $\mathcal{L}_0$. The log-Likelihood function, without systematic effects, is:

$$\begin{aligned}\chi^2 &= -2\ln[\mathcal{L}/\mathcal{L}_0] \\ &= 2\sum_{i=1}^{I}[(t_i - n_i) - n_i \ln(t_i/n_i)]. \end{aligned} \tag{6}$$

One of the virtues of Poisson statistics is the logarithmic behavior with t_i which keeps it positive definite as long as $n_i > 0$.

It is essential to remember, when putting in systematic effects, that the statistical significance in this expression comes from n_i, the number of counts in each bin. To preserve this, the corrections which **multiply the experimental result** in Eq. 4 for Gaussian uncertainties, must here **divide the theoretical prediction**, t_i, i.e.

$$\chi^2 = 2\sum_{i=1}^{I}\left[\left(\frac{t_i}{\prod_k(1+f_i^k S_k)} - n_i\right) - n_i \ln\left(\frac{t_i}{n_i \prod_k(1+f_i^k S_k)}\right)\right] + \sum_k S_k^2 \tag{7}$$

$$= 2\sum_{i=1}^{I}\left[(t_i' - n_i) - n_i \ln\left(\frac{t_i'}{n_i}\right)\right] + \sum_k S_k^2, \tag{8}$$

where the t_i' are shorthand for the theoretical estimate of the number events adjusted by the S_k. In this expression, we continue to treat the constraints, S, on systematic deviations as Gaussian, and their contribution to χ^2 is properly weighted in this expression. Note that some log-likelihood analyses neglect the factor of 2 above, so alertness is required for proper relative weighting of statistical and systematic uncertainties when applying this method to pre-existing log-likelihood code.

3 Backgrounds

Suppose that the event sample is contaminated by a background? Frequently practice is to subtract the background, β_i, from the number of events bin-by-bin, fold in the "Gaussian" uncertainty, and fit the difference to theory. This works adequately with large samples of events, but runs into trouble with low statistics. It also can yield a negative number of events per bin for comparison with theory.

A more correct approach is to let $t_i' \to t_i' + \beta_i$ in Eq. 8. To account properly for the uncertainty in β_i constraints are needed. This depends on how the values were determined.

It they are formulated in a manner similar to the systematic corrections discussed above, then they become just another correlated systematic effect and the formulation is like that above for the S_k. Eq. 8 becomes:

$$\chi^2 = 2\sum_{i=1}^{I}\left[(t_i' + \beta_i(1+S_b f_{bi}) - n_i) - n_i \ln\left(\frac{t_i' + \beta_i(1+S_b f_{bi})}{n_i}\right)\right] + \sum_k S_k^2 + S_b, \quad (9)$$

where the f_{bi} are the 1-std.-dev. limits on the b_i, and S_b is a parameter similar to the S_k.

Another possibility is that the β_i come from some measured number of events, bin-by-bin, in either data or Monte Carlo, properly re-normalized to compare with the signal sample, i.e. $\beta_i = Fb_i$. In this case the backgrounds are statistically independent of each other and obey Poisson statistics. Eq. 8 becomes:

$$\begin{aligned}\chi^2 &= 2\sum_{i=1}^{I}\left[(t_i' + \beta_i - n_i) - n_i \ln\left(\frac{t_i' + \beta_i}{n_i}\right)(\beta_i/F - b_i) - b_i \ln\left(\frac{\beta_i}{Fn_i}\right)\right] \\ &\quad + \sum_k S_k^2 + \left(\frac{F - F_0}{\Delta F}\right)^2, \qquad (10)\end{aligned}$$

where the final term incorporates any uncertainty in the renormalization factor, F.

At first, it seems like the calculation will be enormously more complex. For example, in the analysis of the inclusive jet cross section, there would be 42 data bins and eleven systematics, i.e. 42 β_i's plus ten S_k's plus F. This suggests a 53-parameter fit – possible, but cumbersome. In practice, an enormous simplification can be achieved by noting that, for a given choice of S_k and F, there is an analytic solution for each β_i independent of the others.

$$\begin{aligned}0 &= \frac{\partial\chi^2}{\partial\beta_i} \\ &= 2\left[1 - \frac{n_i}{t_i' + \beta_i} + \frac{1}{F} - \frac{b_i}{\beta_i}\right] \qquad (11)\end{aligned}$$

This is a simple quadratic equation, easily programmed. The quadratic ambiguity is resolved because one sign in the solution is always unphysical. In a χ^2 minimization program like CERNLIB's Minuit[5], each entry to the FCN routine which evaluates χ^2 for a given set of S_k and F can loop over bins and solve exactly for the β_i.

These techniques can be extended in obvious ways for several sources of background computed independently, or for cases of several Monte Carlo samples, each covering some subset of the bins in the signal sample.

4 Efficiencies and Binomial Statistics

Now let us consider a bin-by-bin correction for efficiency. (To keep the equations from becoming cumbersome, we neglect backgrounds, but the discussion is valid when they are included.) To our list of quantities, let us add the following quantities determined independently of the measurement of n_i:

- m_i = number of events in an uncut test sample for measuring the efficiency in the i^{th} bin above.
- k_i = number of events in a subset of the sample which pass selection cuts.
- $e_i = k_i/m_i$ = the efficiency for passing the cuts.

In this treatment, each bin is independent of the others, and we drop the subscript for this part of the discussion. For each event in the sample, there is a probability, e, of passing cuts and a probability, $(1-e)$, of failing cuts. The sum of probabilities is $e+(1-e)=1$. For m events the combined probability can be expressed as $1^m = [e+(1-e)]^m$. We re-write this as a binomial expansion:

$$1 = \sum_k \frac{m!}{k!(m-k)!} e^k (1-e)^{(m-k)} \tag{12}$$

Given the value of e, the probability, $Q_m^k(e)$, of observing k events passing cuts is just the k^{th} term in this expansion:

$$Q_m^k(e) = \frac{m!}{k!(m-k)!} e^k (1-e)^{(m-k)} \tag{13}$$

However, we are usually faced with the oppposite question: given m and k, what is the probability distribution for e? It is proportional to exactly the same term, but its integral must be normalized over the range, $0 \le e \le 1$. This yields:

$$P_m^k(e) = \frac{(m+1)!}{k!(m-k)!} e^k (1-e)^{(m-k)} \tag{14}$$

Now we are in a position to evaluate the mean and variance for e as:

$$\begin{aligned}
\overline{e} &= \frac{k+1}{m+2} && (15)\\
&\approx k/m \qquad \text{(for large } k,\ m) && (16)\\
\overline{(\Delta e)^2} &= \overline{e^2} - \overline{e}^2 &&\\
&= \frac{(k+1)(m-k+1)}{(m+2)^2(m+3)} && (17)\\
&\approx \frac{e(1-e)}{m} \qquad \text{(for large } k,\ m) && (18)
\end{aligned}$$

The *most likely* value for e differs from $\overline{e}$. It is easily evaluated from the derivative of $P_m^k(e)$ to be:

$$e_0 = k/m \tag{19}$$

This seeming discrepancy is a property of asymmetric probability distributions, and becomes more intuitive when one cosiders a measurement which happens to yield $k=0$. The most likely value for e is zero, but non-zero values of e can also yield $k=0$, and the expectation (average) value is $1/(m+2)$. Similar considerations, not discussed above, apply to Poisson statistics.

From Eq. 14, we can form a log-likelihood function:

$$\begin{aligned}
\frac{\mathcal{L}}{\mathcal{L}_0} &= \frac{P_m^k(e)}{P_m^k(e_0)} &&\\
&= \left(\frac{e}{e_0}\right)^k \left(\frac{(1-e)}{(1-e_0)}\right)^{(m-k)} && (20)\\
\chi^2 &= -2\log\left(\frac{\mathcal{L}}{\mathcal{L}_0}\right) &&\\
&= -2\left[k\log\left(\frac{e}{e_0}\right) + (n-k)\log\left(\frac{1-e}{1-e_0}\right)\right] && (21)
\end{aligned}$$

Now we resume the use of subscripts, and modify Eq. 8 to include the efficiency, both as a factor in the Poisson contribution to χ^2 and as an additional binomial contribution to χ^2.

$$\begin{aligned} \chi^2 &= 2\sum_{i=1}^{I}\left[(e_i t'_i - n_i) - n_i \ln\left(\frac{e_i t'_i}{n_i}\right)\right] \\ &\quad + \sum_k S_k^2 \\ &\quad -2\sum_{i=1}^{I}\left[k \log\left(\frac{e_i}{e_{0i}}\right) + (n-k)\log\left(\frac{1-e_i}{1-e_{0i}}\right)\right] \end{aligned} \tag{22}$$

We can now treat the e_i's as a set of parameters to be minimized along with the S_k's. We have not changed the number of degrees of freedom because, for each new parameter in the fit, we have added a constraint in the last term in Eq. 22.

As in the previous section for backgrounds independent in each bin, the bin-by-bin solutions for efficiency are easily determined by:

$$\begin{aligned} 0 &= \frac{\partial \chi^2}{\partial e_i} \\ &= 2\left[t'_i - \frac{n_i + k_i}{e_i} - \frac{m_i - k_i}{1 - e_i}\right] \end{aligned} \tag{23}$$

Again, this is a quadratic equation in e_i, and all the other quantities in it are known. One of the two signs in the solution always yields an unphysical solution.

When the efficiencies are incorporated in this manner, the binomial statistics place a heavy penalty in χ^2 for any comparison between experiment and theory which would require more than 100% efficiency to reconcile data and theory. The Gaussian approximation imposes an unreasonably light penalty for the same situation.

References

[1] **Statistics Issues for the Inclusive Jet E_t Paper** , by J. Huth and S. Behrends, CDF Note CDF/ANAL/JET/CDFR/747, September 5, 1988.

[2] X_t **Scaling (Updated): Comparison of Jet Production at** $\sqrt{s}$ = 546 **GeV and 1800 GeV** , by S. Behrends and A. F. Garfinkel, CDF Note CDF/ANAL/JET/CDFR/1650, February 4, 1992.

[3] **Determining Poisson Errors**, by A. Garfinkel, CDF Note 708, May 26, 1988.

[4] **Notes on Statistics VI. Coping With Low Population Distributions, Maximum Likelihood with Poisson Statistics** , by T. Devlin, November 10, 1988.

[5] **Minuit Reference Manual, Version 94.1** by F. James, CERN Program Library Long Writeup D506, CERN, Geneva, 1994.