Calculating Efficiencies and Their Uncertainties

Marc Paterno
FNAL/CD/CEPA/SLD
paterno@fnal.gov

May 5, 2003

Abstract

The commonly used methods for the calculation of the statistical uncertainties in cut efficiencies ("Poisson" and "binomial" errors) are both defective, as is seen in extreme cases. A method for the calculation of uncertainties based upon Bayes' Theorem is presented; this method has no problem with extreme cases. A program for the calculation of such uncertainties is also available.[1]

1 The Problem

A common technique for the calculation of the efficiency of data selection (a *cut*) makes use of two identically-binned histograms. In histogram **A**, one plots the distribution of the quantity of interest for *all* the events of the sample; in histogram **B** one plots the distribution of the same quantity, but only for those events satisfying the selection criterea, that is, those events which "pass the cut". Intuition leads one to expect that the "best" estimate for the (unknown true) efficiency of the cut for each bin is just k_i/N_i, where N_i is the number of events in bin i of histogram **A**, and k_i is the number of these events which pass the cut – the number of events in bin i of histogram **B**. But what statistical uncertainty should be assigned this estimate of the true efficiency?

There are two commonly used solutions: "Poisson" and "binomial" errors. I believe that neither is correct, because each leads to absurd results in limiting cases. In this paper, I describe these methods, discuss the regimes in which each is appropriate, and discuss the regime

[1] This paper is an updated version of a 1996 paper of the same name. The mathematical content is the same, but the exposition has been slightly expanded, the author's contact information has been updated, and information on the new software available has been added.

in which each breaks down. I then present a simple alternative, not subject to these defects.

For simplicity of notation, from here forward let us consider only a single bin, in which k events out of a total of N events pass the cuts. To determine the uncertainties in each bin of a histogram, one merely applies the same rule independently to each bin i. Note that this method calculates *independent* statistical uncertainties for each bin; in this it is similar to both of the commonly used methods.

2 Two Common Procedures

2.1 Poisson Errors

2.1.1 Calculational Technique

In the “Poisson” calculation, one applies the large sample limit of the Poisson distribution, to say that the uncertainty δ_k in k is $\sqrt{k}$, and that the uncertainty δ_N in N is $\sqrt{N}$. Then, using the standard error propagation formula[2], one finds

$$\begin{aligned}
\delta\epsilon' &= \epsilon'\sqrt{\left(\frac{\delta k}{k}\right)^2 + \left(\frac{\delta N}{N}\right)^2} \\
&= \frac{k}{N}\sqrt{\frac{1}{k} + \frac{1}{N}} \\
&= \sqrt{\frac{k^2(N+k)}{N^3}}. \qquad (1)
\end{aligned}$$

Among the arguments against this method is the behavior in limiting cases.

2.1.2 Examples of Failure

As a first example, consider the case $k = 0$, for any $N \geq 1$, we find $\epsilon' \pm \delta\epsilon' = 0 \pm 0$. The calculation is telling us that if we observe one event, and it fails the cut, we know *with complete certainty* that the efficiency is exactly zero. This is a remarkable conclusion, which differs greatly from our intuition.

As a second example, consider the case $k = N$. Then we find $\epsilon' \pm \delta\epsilon' = 1 \pm \sqrt{2/N}$. So if we have a sample with a single event, and if the event passes our cuts, then we find $\epsilon' \pm \delta\epsilon' = 1 \pm 1.4$; the “$1\sigma$” error interval spans the range $[-0.6, 2.4]$. Again this is a remarkable result — most of this range is unphysical! In fact, when we have $k = N$,

[2] For a standard justification of the standard formula, see, for example, *Data Reduction and Error Analysis for the Physical Sciences*, Phillip R. Bevington.

for *any* value of N, we find that the error range extends past 1, even though by definition it is impossible that we should have $\epsilon > 1$. Clearly the "Poisson" error calculation is in disagreement with our reasonable expectations.

2.1.3 Regime of Validity

Of course, few would try to use "Poisson" uncertainties in the small-k limit; the approximation of $\sqrt{k}$ as the uncertainty in an observation of k events only holds for large k. In such small-k cases, some expect better behavior from the "binomial" errors, presented in the next session. However, the failure of the case where N is large, and k is near N, is not due to a failure of the "small N" limit; it is an inherent failure of the model.

2.2 Binomial Errors

2.2.1 Calculational Technique

Next let us consider the "binomial" error calculation. This calculation is based on the knowledge that the application of a cut (or cuts) can be considered a binomial process, with probability of "success" (*i.e.* true efficiency) ϵ.

Given the true efficiency ϵ and the sample size N, the expectation value for the number of events $\langle k \rangle$ passing the cut is given by $\langle k \rangle = \epsilon N$, and the standard deviation of the distribution of the number of events passing is

$$\begin{aligned} \sigma_k &= \sqrt{\mathrm{var}(k)} \\ &= \sqrt{\epsilon(1-\epsilon)N}. \end{aligned} \tag{2}$$

Since we don't know the true efficiency, what is often done is to put our estimate ϵ' into this equation in its place, and then to divide through by N, yielding the result $\delta\epsilon' = (1/N)\sqrt{k(1-k/N)}$.

2.2.2 Examples of Failure

While the "binomial errors" equation doesn't contain the absurdity of error ranges extending into unphysical regions below zero or above one, it still yields absurd results in limiting cases. Consider, for example, what we find when we observe only a single event. If it passes, we find $\epsilon' \pm \delta\epsilon' = 1 \pm 0$; if it fails, we find $\epsilon' \pm \delta\epsilon' = 0 \pm 0$. In each case, this calculation claims perfect certainty for the measured efficiency. In fact, for all N, if we have either $k = 0$ or $k = N$, then we find a zero error. Again, this violates our reasonable expectation.

2.2.3 Region of Validity

The "binomial errors" approximation works acceptably when k is neither "too close" to 0, nor "too close" to N. The determination of what is "too close" is a matter of judgement; clearly both $k = 0$ and $k = N$ are too close.

In the next section, I develop a calculation, based on the use of Bayes' Theorem, that calculates the statistical uncertainty in the efficiency in a manner that agrees with our reasonable expectations, and which exhibits reasonable behavior even in limiting cases.

3 A Better Calculation

Let us start out again with the binomial probability, this time writing it more explicitly. The symbol $P(k|\epsilon, N, I)$ denotes the probability that k events will pass the cut, given the conditions that the true efficiency is ϵ, that there are N events in the sample, and that our prior information I tells us this is a binomial process. The binomial distribution is

$$P(k|\epsilon, N, I) = \frac{N!}{k!(N-k)!}\epsilon^k(1-\epsilon)^{N-k}. \tag{3}$$

In our problem, we do not know ϵ; rather, we have our data, which is an observation of k events out of N passing the cut. What we need to determine then is $P(\epsilon|k, N, I)$, which is the probability that the true efficiency is between ϵ and $\epsilon + d\epsilon$. Knowing this, we can determine the most probable value of ϵ (that is, the value of ϵ we determine to be most probable, given our data) and also a confidence interval for ϵ with known probability content, so that we may make comparisons with some statistical meaning. So how do we find $P(\epsilon|k, N, I)$?

3.1 Bayes' Theorem

The method for inverting a probability (determining $P(A|BC)$ from $P(B|AC)$) is called Bayes' Theorem. In this context, it is

$$P(\epsilon|k, N, I) = \frac{P(k|\epsilon, N, I)P(\epsilon|N, I)}{\mathcal{Z}}, \tag{4}$$

where $\mathcal{Z}$ is a constant to be determined by normalization, and $P(\epsilon|N, I)$ is the probability we assign for the true efficiency to be between ϵ and $\epsilon + d\epsilon$ *before* we consider the data. What shall we assign for this probability? A moment's consideration tells us that, given only N and the fact that we are dealing with a binomial process, so that ϵ must be in the inclusive range $[0, 1]$, we would have no reason to favor one value of the efficiency over another. Therefore it is reasonable to take

$$P(\epsilon|N,I) = \begin{cases} 1 & \text{if } 0 \le \epsilon \le 1 \\ 0 & \text{otherwise} \end{cases} , \tag{5}$$

independent of N. Note that the use of probability theory allows us to include in our calculation the knowledge that the efficiency must be between zero and one; this knowledge is built into the pre-data probability distribution describing our knowledge of ϵ, which assigns zero probability to those values of ϵ which we know, with certainty, to be impossible.

With $P(k|\epsilon, N, I)$ given by (3), we have all that is needed to determine the normalization constant $\mathcal{Z}$ and thus the probability distribution describing post-data knowledge of ϵ.

To determine the normalization constant $\mathcal{Z}$, we must solve

$$\begin{aligned} &\int_{-\infty}^{\infty} P(\epsilon|k,N,I)\,d\epsilon \\ &\quad = \frac{\int_{-\infty}^{\infty} P(k|\epsilon,N,I)P(\epsilon|N,I)\,d\epsilon}{\mathcal{Z}} \\ &\quad = \frac{1}{\mathcal{Z}}\frac{N!}{k!(N-k)!}\int_0^1 \epsilon^k(1-\epsilon)^{N-k}\,d\epsilon \\ &\quad = 1 \end{aligned} \tag{6}$$

for $\mathcal{Z}$. Noting that the Beta function $B(x,y)$ is defined by $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$, we may determine the normalization constant. We use the relation $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$, and find

$$P(\epsilon|k,N,I) = \frac{\Gamma(N+2)}{\Gamma(k+1)\Gamma(N-k+1)}\epsilon^k(1-\epsilon)^{N-k}. \tag{7}$$

3.2 Using the Solution

Figure 1 shows $P(\epsilon|k,N,I)$ for $N = 5$ and $k = 0, 1, 2, 3, 4$ and 5. Note that in all cases, we assign zero probability that ϵ is below zero or above one. Note also that we assign zero probability to $\epsilon = 0$ unless $k = 0$; this is necessary, of course, since if we observe even a single event which passes, we know the efficiency cannot be zero. Similarly, we assign zero probability to $\epsilon = 1$ unless $k = N$, since if even a single event fails our cut, we know that the efficiency is not one. Finally, note that in each case the peak of the distribution — and thus the most probable value for the efficiency, given the data — is at exactly k/N. Thus our intuition leads us to the same result as does the application of probability theory.

To verify the observation of the location of the peak of the probability, we find the maximum of $P(\epsilon|k,N,I)$ by solving

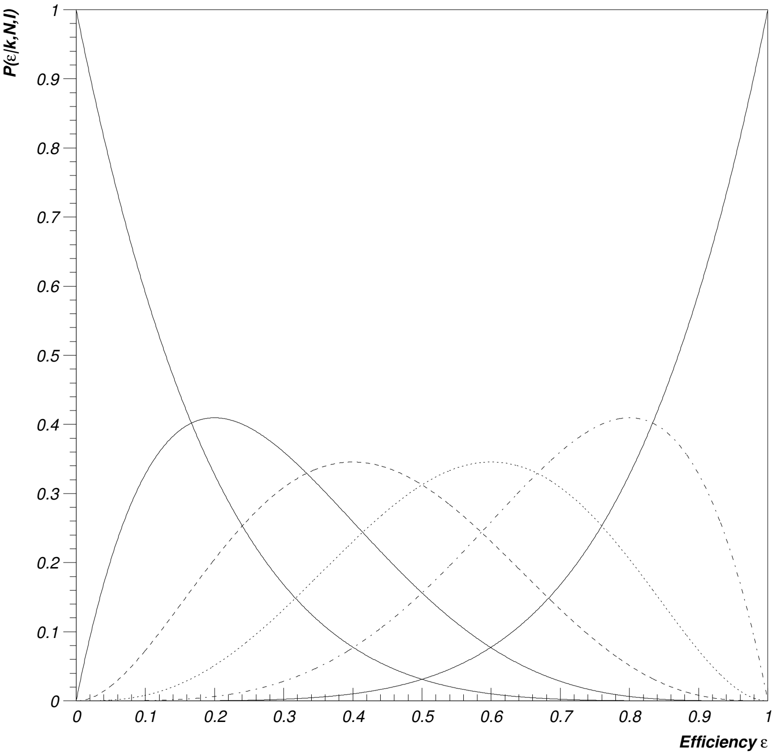

Figure 1: The post-data probability distribution for the cut efficiency ϵ, for $N = 5$, and for $k = 0, 1, 2, 3, 4$ and 5.

$$\frac{dP(\epsilon|k,N,I)}{d\epsilon} = \frac{\Gamma(N+2)\epsilon^k(1-\epsilon)^{N-k}(N\epsilon - k)}{\Gamma(k+1)\Gamma(N-k+1)\epsilon(\epsilon-1)}. \tag{8}$$

The solutions for the extrema are $\epsilon = 0$, $\epsilon = 1$, and $\epsilon = k/N$. Investigation of the second derivative (or inspection of the plot in Figure 1) shows that the extrema at $\epsilon = 0$ and $\epsilon = 1$ are minima, except when $k = 0$ or $k = N$. In all cases, $\epsilon = k/N$ is a maximum.

Since the probability distribution describing our knowledge of ϵ is not a delta function, our estimate of the true efficiency has some uncertainty. There are many measures of this uncertainty that can be extracted from the distribution $P(\epsilon|k,N,I)$: upper and lower limits at various confidence levels; the variance, or its square root, the standard deviation; the mean absolute deviation; or confidence intervals of various sorts.

I recommend using the *shortest 68.3% confidence interval* as the measure of the uncertainty in the efficiency measurement. It has two attractive features. First, it has a known probability content, one chosen to be the same as a “$1\ \sigma$” Gaussian error. Therefore such error intervals will behave as we most often expect. Second, it is the most constrained region which has this probability content, so that we present our measurement in the fashion that most constrains the range in which we believe the true value exists. It has the unfortunate drawback of being challenging to calculate. Unlike the formula for the peak of the distribution, the formula for the shortest interval containing probability content λ is non-trivial. To find the shortest interval $[\alpha, \beta]$ which contains probability content λ, we must minimize the interval $\beta - \alpha$, subject to the constraint

$$\int_{\alpha}^{\beta} P(\epsilon|k, N, I)\, d\epsilon = \lambda. \tag{9}$$

A formal solution can be found using the method of Lagrange multipliers. The solution for α and β is found by the simultaneous solutions of the nonlinear equations[3]

$$\begin{aligned} \mathcal{G} + \rho\alpha^k(1-\alpha)^{N-k} &= 0 \\ \mathcal{G} + \rho\beta^k(1-\beta)^{N-k} &= 0 \\ \mathrm{B}_\beta(k+1, N-k+1) - \mathrm{B}_\alpha(k+1, N-k+1) &= \lambda\mathcal{G} \end{aligned} \tag{10}$$

where λ is the required probability content of the confidence interval (for example, 0.683), $\mathcal{G} = \Gamma(k+1)\Gamma(N-k+1)/\Gamma(N+2)$, and ρ is a Lagrange multiplier introduced for the constraint. Here $\mathrm{B}_x(u, v)$ is the *incomplete Beta function*, defined by

$$\mathrm{B}_x(u, v) = \int_0^x t^{u-1}(1-t)^{v-1}\, dt. \tag{11}$$

3.3 Software

Since no closed-form solution of these equations is at hand (at least, not one that I can find), and since the numerical solution of simultaneous nonlinear equations is a difficult problem, it turns out to be simpler to write a program which minimizes $\beta - \alpha$ subject to the integral constraint directly. I have made available CALCEFF, a program that does just that. This program takes as input a file of pairs of k and N, and and a probability content. The output of the program is a table containing the most probable efficiency k/N, and α and β, the lower

[3] Readers of the 1996 version of the paper may note there is a change in notation from that version. The mathematical content is the same.

and upper edges of the shortest confidence interval with the required probability content.

A simple library routine which performs the calculation for a single point, as well as the program CALCEFF, are both available in source format from the author. The source code is standard C++, and should be handled by any reasonably conformant compiler. Please contact the author at paterno@fnal.gov to obtain a copy of the code.