CDF/MEMO/STATISTICS/PUBLIC/7168
Version 1.0

# Uncertainties on Efficiencies

Craig Blocker
Brandeis University
August 4, 2004

**Abstract**

A brief discussion of uncertainties on efficiencies is given. Four cases are considered: (1) the simplest counting case, (2) a case where the number of possible events and number of successful events are determined by fits to distributions, (3) a case where the number of events is determined by side band subtraction of a distribution, and (4) a case with side-band subtracted, weighted events.

## 1 Introduction

Recently, a couple of questions have been submitted to the Statistics Committee concerning uncertainties on efficiencies. Both questions were of the same nature, namely, how do you deal with the correlation between the number of events that pass a cut with the total number before the cut. The Statistics Committee has developed formulae to cover the cases in question. These are of sufficient generality to warrant a note. In addition, the method is presented so that people can develop their own formula if their case is different.

One important point is that testing is essential, in all the but the simplest cases. All of the formulae in this note were tested on large samples of simple Monte Carlo data and found to correctly predict the spread in measured efficiencies. It is not possible to test all possible values of parameters, but for some reasonable choices, the formulae here work very well. The reader is urged to test these formulae in their particular case before using them.

Section 2 covers a very simple counting case with the well-known binomial result. An alternative derivation is presented that forms the basis for later derivations. Section 3 discusses the case where the number of events before and after the cut are determined by fits to a distribution. Section 4 covers the case where a side band subtraction is done, and Section 5 extends the side band method to weighted histograms.

## 2 Simple Counting

The simplest case is just counting the number $N_0$ of candidate events and the number $N_p$ that pass a cut. The efficiency is then given by

$$\epsilon = \frac{N_p}{N_0}. \tag{1}$$

Since $N_p$ and $N_0$ are correlated, using equation 1 with propagation of uncorrelated Poisson errors does not give the correct uncertainty on the efficiency. Usually, this is handled by noting that this is equivalent to a binomial problem with total events $N_0$ and a probability $\epsilon$ for each event to pass. The uncertainty on $\epsilon$ is then given by

$$(\Delta\epsilon)^2 = \frac{\epsilon(1-\epsilon)}{N_0}. \tag{2}$$

An equivalent, alternative method is to consider the number $N_p$ of events that pass and the number $N_f$ that fail (see pages 46-48 of **Statistics for Nuclear and Particle Physicists** by Louis Lyons, Cambridge University Press, 1986). These two are uncorrelated and hence easier to use in error propagation. Note that in this approach, the total number of events $N_0 = N_p + N_f$ is not a fixed number, but is itself Poisson distributed. The efficiency is

$$\epsilon = \frac{N_p}{N_p + N_f}. \tag{3}$$

Standard error propagation then gives

$$\begin{aligned}
(\Delta\epsilon)^2 &= \left(\frac{\partial\epsilon}{\partial N_p}\right)^2 (\Delta N_p)^2 + \left(\frac{\partial\epsilon}{\partial N_f}\right)^2 (\Delta N_f)^2 && (4)\\
&= \left(\frac{N_f}{N_0^2}\right)^2 (\Delta N_p)^2 + \left(\frac{-N_p}{N_0^2}\right)^2 (\Delta N_f)^2 && (5)\\
&= \frac{(1-\epsilon)^2 N_p + \epsilon^2 N_f}{N_0^2} && (6)\\
&= \frac{\epsilon(1-\epsilon)}{N_0}. && (7)
\end{aligned}$$

Note that this is exactly the same result as obtained by considering it as a binomial problem, as it should be since they are equivalnet. The reason for considering the second method is that it is easier to extend to the cases considered below.

Also note that in practice you don't know the true $\epsilon$ and use the measured value from equation 1 or 3. If $\epsilon$ is close to 0 or 1, this is usually not a good approximation, since $(\Delta\epsilon)^2$ varies relatively rapidly with $\epsilon$ in this case. Handling this effect is not the subject of this note (see CDF note 5894 by John Conway for a Bayesian treatment of this subject).

## 3 Efficiency From Fits

Often there is background to the events of interest, and fits are done to determine the numbers. For example, we might be interested in the number of $J/\psi$'s before and after some cut. We fit to the mass distribution of the $\mu^+\mu^-$ including a background term to determine the number of signal events. The number before the cut $N_0$ is correlated with the number $N_p$ after the cut, so simple error propagation in these variables is not feasible.

Instead, suppose that the fit number that pass the cut is $N_p \pm \Delta N_p$ and the fit number that fail the cut is $N_f \pm \Delta N_f$. The efficiency is

$$\epsilon = \frac{N_p}{N_p + N_f}. \tag{8}$$

Standard error propagation gives

$$\begin{aligned}
(\Delta\epsilon)^2 &= \left(\frac{\partial\epsilon}{\partial N_p}\right)^2 (\Delta N_p)^2 + \left(\frac{\partial\epsilon}{\partial N_f}\right)^2 (\Delta N_f)^2 && (9)\\
&= \frac{(1-\epsilon)^2(\Delta N_p)^2 + \epsilon^2(\Delta N_f)^2}{N_0^2}, && (10)\\
& && (11)
\end{aligned}$$

where we assume $N_0 = N_p + N_f$ (which is not exactly true in each case since each of these numbers comes from a fit, but is a hopefully good approximation).

If we also assume that $(\Delta N_0)^2 = (\Delta N_p)^2 + (\Delta N_f)^2$, then we can rewrite $(\Delta\epsilon)^2$ completely in terms of results of fits to the total number before the cut and the number that pass the cut, that is,

$$\begin{aligned}
(\Delta\epsilon)^2 &= \frac{(1-\epsilon)^2(\Delta N_p)^2 + \epsilon^2(\Delta N_f)^2}{N_0^2} && (12)\\
&= \frac{(1-2\epsilon)(\Delta N_p)^2 + \epsilon^2((\Delta N_p)^2 + (\Delta N_f)^2)}{N_0^2} && (13)\\
&= \frac{(1-2\epsilon)(\Delta N_p)^2 + \epsilon^2(\Delta N_0)^2}{N_0^2}. && (14)
\end{aligned}$$

Note that if we replace $(\Delta N_p)^2$ and $(\Delta N_0)^2$ by their Poisson values of $N_p$ and $N_0$, respectively, we get back the usual binomial formula.

It is important to note that in the equations above, uncertainties on numbers are the variations we would expect if we repeated the measurements and got variations in both the signal fraction and the total number of events. Often, a distribution containing signal and background contributions is fit with a parameter $f$ giving the fraction of signal. If the uncertainty from the fit on $f$ is $\Delta f$ and the total number of events being fit is $N_{tot}$, then the

number of signal events $N_s$ and its uncertainty are given by

$$\begin{aligned} N_s &= fN_{tot} && (15)\\ (\Delta N_s)^2 &= (\Delta f)^2 N_{tot}^2 + f^2(\Delta N_{tot})^2 && (16)\\ &= (\Delta f)^2 N_{tot}^2 + f^2 N_{tot}. && (17) \end{aligned}$$

This uncertainty can also be obtained by doing an extended likelihood fit, where a Poisson term for $N_{tot}$ is included in the likelihood function (see pages 98-100 of **Statistics for Nuclear and Particle Physicists** by Louis Lyons, Cambridge University Press, 1986).

## 4 Efficiency From Side Band Subtraction

Another technique for determining an efficiency from a distribution with background is side band subtraction. For example, we may have a sample of $J/\psi$'s and determine the numbers before and after a cut by a side band subtraction in the mass distributions.

We define a signal region and a side band region. Let $N_p$ and $N_f$, be the numbers of events in the signal region that pass and fail the cut, respectively. Let $N_{p,SB}$ and $N_{f,SB}$ be the corresponding numbers in the side bands. Define $N_0 = N_p + N_f$ and $N_{0,SB} = N_{p,SB} + N_{f,SB}$. We want to include the fact that the side bands may not have the same number of expected background events as the signal region by defining the ratio of expected events to be $\alpha$, that is, if there are $N_{SB}$ side band events, we expect $\alpha N_{SB}$ events in the signal region. In this derivation, it is assumed that $\alpha$ is the same before and after the cut. If this is not the case, the reader is left to extend the derivation.

The efficiency is

$$\epsilon = \frac{N_p - \alpha N_{p,SB}}{N_p + N_f - \alpha(N_{p,SB} + N_{f,SB})} \qquad (18)$$

Standard propagation of errors gives

$$\begin{aligned} (\Delta\epsilon)^2 &= \left(\frac{\partial\epsilon}{\partial N_p}\right)^2 (\Delta N_p)^2 + \left(\frac{\partial\epsilon}{\partial N_{p,SB}}\right)^2 (\Delta N_{p,SB})^2 + \\ &\quad \left(\frac{\partial\epsilon}{\partial N_f}\right)^2 (\Delta N_f)^2 + \left(\frac{\partial\epsilon}{\partial N_{f,SB}}\right)^2 (\Delta N_{f,SB})^2 && (19)\\ &= \frac{(1-\epsilon)^2((\Delta N_p)^2 + \alpha^2(\Delta N_{p,SB})^2) + \epsilon^2((\Delta N_f)^2 + \alpha^2(\Delta N_{f,SB})^2)}{(N_0 - \alpha N_{0,SB})^2} && (20)\\ &= \big[(1-2\epsilon)((\Delta N_p)^2 + \alpha^2(\Delta N_{p,SB})^2) + \epsilon^2((\Delta N_p)^2 + (\Delta N_f)^2) + \\ &\quad \epsilon^2\alpha^2((\Delta N_{p,SB})^2 + (\Delta N_{f,SB})^2)\big] / (N_0 - \alpha N_{0,SB})^2 && (21)\\ &= \frac{(1-2\epsilon)(N_p + \alpha^2 N_{p,SB}) + \epsilon^2(N_0 + \alpha^2 N_{0,SB})}{(N_0 - \alpha N_{0,SB})^2}, && (22) \end{aligned}$$

where the first line again involves no cross terms because the regions are indepedent and the last step uses the fact that we are counting the number of events in each region, and hence the uncertainties on these number is given by the standard Poisson values. Note that the last form depends only on the numbers that pass and the total numbers (not on the number that fail).

## 5 Efficiency From Weighted Side Band Subtraction

Consider the case where we wish to do a side band subtraction on a weighted histogram. The question that was put to the Statistics Committee concerned weighting only the distribution of events that pass the cut. Specifically, it was a case where the efficiency had a $P_T$ dependence that had been measured and it was desired to look for dependence on other variables by weighting by $1/\epsilon(P_T)$. The "weighted efficiency" is the ratio of the side band subtracted events in the weighted distribution of events that pass to the side band subtracted number of events in the unweighted total.

Define $N_p$, $N_f$, $N_{p,SB}$, $N_{f,SB}$, $N_0$, $N_{0,SB}$, and $\alpha$ as in section 4. Let $w(x)$ be the weight that depends on some external variable (for example, $x = P_T$ in the example above). The weighted sums are

$$W_p = \sum_{i=1}^{N_p} w(x_i) \tag{23}$$

$$W_{p,SB} = \sum_{i=1}^{N_{p,SB}} w(x_i), \tag{24}$$

where is the sum is over the events in the appropriate region. The "weighted efficiency" is defined as

$$\epsilon = \frac{W_p - \alpha W_{p,SB}}{N_0 - \alpha N_{0,SB}}. \tag{25}$$

When we propagate the errors, we need the uncertainty on $W_p$. This has two contributions - one from variation in the sampling of $x$ and one from variation in the number of events. First, consider the variation due to sampling of $x$, that is, we could repeat the experiment and get the same number $N_p$ of events, but get a different $W_p$ because the set of weights is different. In this case,

$$\overline{W_p} = \overline{\sum_{i}^{N_p} w(x_i)} \tag{26}$$

$$= N_p \overline{w} \tag{27}$$

$$\overline{W_p^2} = \overline{\sum_{i}^{N_p} w(x_i) \sum_{j}^{N_p} w(x_j)} \tag{28}$$

$$= \overline{\sum_{i}^{N_p} w^2} + \overline{\sum_{i \neq j} w_i w_j} \quad (29)$$

$$= N_p \overline{w^2} + (N_p^2 - N_p)\overline{w}^2 \quad (30)$$

$$(\Delta W_p)^2 = \overline{W_p^2} - \overline{W_p}^2 \quad (31)$$

$$= N_p \Delta w^2 \quad (32)$$

The second contribution is from variation in the number of events that pass the cut. We can approximate this as

$$\Delta W_p = \sum_{i=N_p+1}^{N_p+\Delta N_p} \approx \Delta N_p \overline{w}. \quad (33)$$

This latter term is clearly correlated with $\Delta N_p$.

There are similar contributions to the uncertainty on $W_{p,SB}$. We allow for the possibility that the averages on $w$ are different for the background and the signal, giving

$$(\Delta W_{p,SB})^2 = N_{p,SB}(\Delta w_{SB})^2 + (\Delta N_{p,SB})^2 \overline{w}_{SB}^2, \quad (34)$$

where the first term is uncorrelated with $\Delta N_{p,SB}$ and the second term is fully correlated.

Propagating the errors on equation 25, including the correlations gives

$$(\Delta\epsilon)^2 = \left[(\overline{w}-\epsilon)^2(\Delta N_p)^2 + \alpha^2(\overline{w}_{SB}-\epsilon)^2(\Delta N_{p,SB})^2 + \epsilon^2((\Delta N_f)^2 + \alpha^2(\Delta N_{f,SB})^2) + N_p(\Delta w)^2 + \alpha^2 N_{p,SB}^2\right]/(N_0 - \alpha N_{0,SB})^2 \quad (35)$$

$$= \left[(\overline{w}^2 - 2\overline{w}\epsilon)N_p + (\overline{w}_{SB}^2 - 2\overline{w}_{SB}\epsilon)\alpha^2 N_{p,SB} + \epsilon^2(N_0 + \alpha^2 N_{0,SB}) + N_p(\Delta w)^2 + \alpha^2 N_{p,SB}\right]/(N_0 - \alpha N_{0,SB})^2, \quad (36)$$

where Poisson uncertainties have been used in the last step. Note that these formulae only consider variations in $w$ due to its dependence on the external variable $x$. If there is additional significant uncertainty, for example, a statistical uncertainty on the measured weight function, it would also need to be included.

The unweighted case in section 4 is a special case of the weighted case, as can be seen by setting the average weights to 1 and the $\Delta w$'s to 0 in equation 36, which yields equation 22.

## 6 Acknowedgements

I would like to thank the CDF Statistics Committee for many useful discussions about the issues in this note.