# Treatment of Errors in Efficiency Calculations

T. Ullrich and Z. Xu
Brookhaven National Laboratory

January 18, 2007

**Abstract**

In this report we discuss the treatment of statistical errors in cut efficiencies. The two commonly used methods for the calculation of the errors, Poissonian and Binomial, are shown to be defective. We derive the form of the underlying probability density function and characterize its mean, mode, and variance. A method for the calculation of errors based on the variance of the distribution is discussed.

## 1 Introduction

In many areas of experimental particle and nuclear physics the efficiencies of the detectors used to record the particles is evaluated by Monte Carlo simulations. These simulations usually incorporate a generator that produces particles and events with known parameters and a detailed simulation of the detectors and their respective response. The produced Monte Carlo data is then run through the standard event reconstruction software and the result is compared to the input from the generator. The ratio of output over input defines the so called response function of the detector. This function is then used to correct the result obtained from the reconstruction of the real data yielding the final corrected spectra. The response function depends in general on many parameters, *e.g.*, transverse momentum and rapidity of the particle, event multiplicity and much more.

From a statistics point of view this procedure can be simplified to the comparison of two histograms. In histogram A, one plots the distribution of the quantity of interest for all the data of the sample; in histogram B one plots the distribution of the same quantity, but only for those satisfying the selection criteria, *i.e.*, those data that pass the cuts. 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 $k_i$ is the number of entries in bin $i$ of histogram B and $n_i$ is the number of entries in bin $i$ of histogram A.

But what uncertainty should be assigned to each efficiency? 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 errors in a histogram, one merely applies the same rule independently to each bin $i$. Please note that in the following we consider only statistical errors and ignore all aspects of systematic errors.

# 2 Common but incorrect error evaluations

There are two frequently used, but incorrect, approaches to evaluate the errors of the efficiency calculation. They are briefly discussed below, before we turn to the derivation of the correct treatment of the uncertainties.

## 2.1 Poissonian errors

Assuming that $k$ and $n$ fluctuate independently, the Poisson distribution in the large sample limit tells us that the error $\sigma_k$ in $k$, the output sample, is $\sqrt{k}$, and that the error $\sigma_n$ in $n$, the input sample, is $\sqrt{n}$. Then, using standard error propagation, and given the usual estimator for the real (unknown) efficiency

$$E(\varepsilon) = \hat{\varepsilon} = \frac{k}{n}. \tag{1}$$

one finds for the variance:

$$\mathrm{V}(\hat{\varepsilon}) = \sigma_{\hat{\varepsilon}}^2 = \hat{\varepsilon}^2 \left( \frac{1}{k} + \frac{1}{n} \right) \tag{2}$$

This calculation is flawed for various reasons:

- $n$ is a fixed quantity and not subject to any fluctuation. It's usually a well defined and known input quantity. This reduces Eq. 2 to

$$\mathrm{V}(\hat{\varepsilon}) = \frac{\hat{\varepsilon}^2}{k} = \frac{k}{n^2}. \tag{3}$$

- Furthermore $n$ and $k$ are not independent but highly correlated. Statistics tells us that $k$ must be distributed according to a Binomial probability density function.

- Another strong argument against this method is the behavior in limiting cases. In the case $k = 0$ and $n \geq 1$ one finds $\hat{\varepsilon} = 0$ and $\sigma_{\hat{\varepsilon}} = 0$. The calculation is telling us that if we observe one event, and it fails the cut, we know with complete certainty (zero error) that the efficiency is exactly zero. This is a remarkable conclusion, which differs greatly from our intuition.

Clearly equation 3 (and even more so 2) are in disagreement with our reasonable expectations and basic statistics laws. We conclude that the Poissonian error evaluation is incorrect and is not applicable in our case.

### 2.2 Binomial errors

Next let us consider a simple 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" $\varepsilon$, the *true* efficiency. Given this efficiency and the sample size $n$, the number of events passing the cut is given by a Binomial distribution

$$P(k;\varepsilon,n) = \binom{n}{k}\varepsilon^k(1-\varepsilon)^{n-k} \tag{4}$$

with mean $\overline{k} = \varepsilon n$, and variance $\mathrm{V}(k) = \sigma_k^2 = n\varepsilon(1-\varepsilon)$. Using again the usual estimator for $\hat{\varepsilon} = k/n$, where $n$ in our case is a given input constant we can write down the variance for $\hat{\varepsilon}$ using simple error propagation:

$$\mathrm{V}(\hat{\varepsilon}) = \sigma_{\hat{\varepsilon}}^2 = \frac{\hat{\varepsilon}(1-\hat{\varepsilon})}{n} = \frac{k(n-k)}{n^3} \tag{5}$$

However, this equation also yields absurd results in limiting cases.

1. In the case $k = 0$ we obtain an unphysical zero error as is the case in the Poissonian error calculation (Eq. 3).
2. In the other limit, $k = n$, the formula again yields zero error.

In each case, this calculation claims perfect certainty for the measured efficiency. Again, this violates our reasonable expectation.

In the next section, we now 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 intuition, and that exhibits reasonable behavior even in limiting cases.

## 3 Correct treatment of errors

We start out again with the Binomial probability. $P(k;\varepsilon,n)$ defined in Eq. 4 denotes the probability that $k$ events will pass the cut, given the conditions that the *true* efficiency is $\varepsilon$, that there are $n$ events in the sample, and that our prior information tells us this is a Binomial process.

In our problem, we do not know $\varepsilon$; rather, we have our data, which is an observation of $k$ events out of $n$ passing the cut. What we need to determine is the probability density function $P(\varepsilon;k,n)$, which gives the probability function of $\varepsilon$ for a given $n$ and $k$. Once known, we can determine easily the mean, variance, most probable value, and confidence intervals, so that we can make comparisons with some statistical meaning.

### 3.1 Derivation of the probability density function

In order to calculate we use the Bayesian theorem and make the following ansatz:

$$P(\varepsilon; k, n) = \frac{P(k; \varepsilon, n)\ P(\varepsilon; n)}{C} \tag{6}$$

where $C$ is a constant to be determined by normalization, and $P(\varepsilon; n)$ is the probability we assign for the true efficiency before we consider the data. Given only $n$ and the fact that we are dealing with a Binomial process says simply that $\varepsilon$ must be in the inclusive range $0 \leq \varepsilon \leq 1$; we would have no reason to favor one value of the efficiency over another. Therefore it is reasonable to take

$$P(\varepsilon; n) = \begin{cases} 1 & \text{if } 0 \leq \varepsilon \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

independent of $n$. 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 predata probability distribution describing our knowledge of $\varepsilon$, which assigns zero probability to those values of $\varepsilon$ which we know, with certainty, to be impossible.

To determine the normalization we must solve

$$\int_{-\infty}^{+\infty} P(\varepsilon; k, n)\, d\varepsilon = \frac{1}{C}\binom{n}{k} \int_0^1 \varepsilon^k (1-\varepsilon)^{n-k} d\varepsilon = 1 \tag{8}$$

for $C$. For the calculation of the integral it is useful to recall the definition of the Beta function:

$$B(\alpha+1, \beta+1) = \int_0^1 x^\alpha (1-x)^\beta dx = \frac{\Gamma(\alpha+1)\,\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}. \tag{9}$$

Note also the trivial relation for integer values $\Gamma(n+1) = n!$. Thus we directly obtain

$$C = \frac{n!}{(n+1)!} = \frac{1}{n+1} \tag{10}$$

and the final efficiency probability density function thus reads:

$$P(\varepsilon; k, n) = (n+1)\binom{n}{k}\varepsilon^k (1-\varepsilon)^{n-k} \tag{11}$$

$$= \frac{(n+1)!}{k!\,(n-k)!}\,\varepsilon^k (1-\varepsilon)^{n-k} \tag{12}$$

Figure 1 shows $P(\varepsilon; k, n)$ for $n = 10$ and $k = 0, 1, \ldots, 10$. Note that in all cases, we assign zero probability that $\varepsilon$ is below zero or above one. Note also that we assign zero probability to $\varepsilon = 0$ 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 $\varepsilon = 1$ unless $k = n$, since if even a single event fails our cut, we know that the efficiency is not one.

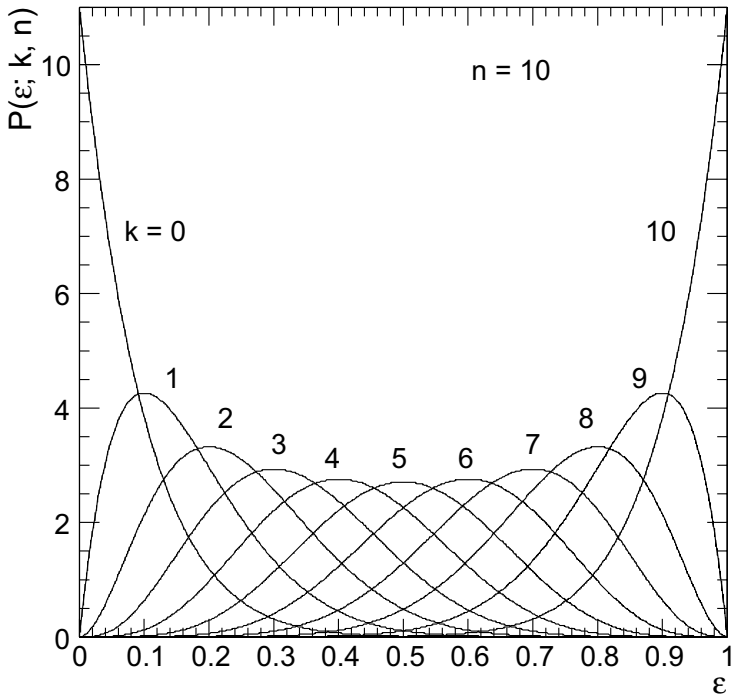


Figure 1: The probability density function $P(\varepsilon; k, n)$ for $n = 10$ and $k = 0, 1, \ldots, 10$.

### 3.2 Features of the probability density function

Given the analytic form we now can calculate the moments of the distribution. For the mean we obtain, again using Eq. 9 to solve the integral:

$$\begin{aligned} \overline{\varepsilon} &= \int_0^1 \varepsilon \, P(\varepsilon; k, n) \, d\varepsilon & (13) \\ &= \frac{(n+1)!}{k!\,(n-k)!} \int_0^1 \varepsilon^{k+1} (1-\varepsilon)^{n-k} d\varepsilon & (14) \\ &= \frac{k+1}{n+2} & (15) \end{aligned}$$

The most probable value (the mode) mode($\varepsilon$) can be easily calculated by solving $dP/d\varepsilon = 0$. We get:

$$\text{mode}(\varepsilon) = \frac{k}{n}. \tag{16}$$

This is a remarkable result. The common estimator of the real efficiency $\hat{\varepsilon} = k/n$ actually is the mode and not the mean, *i.e.*, the expectation value of the distribution. The mean and the mode only become identical for $n = 2k$, and of

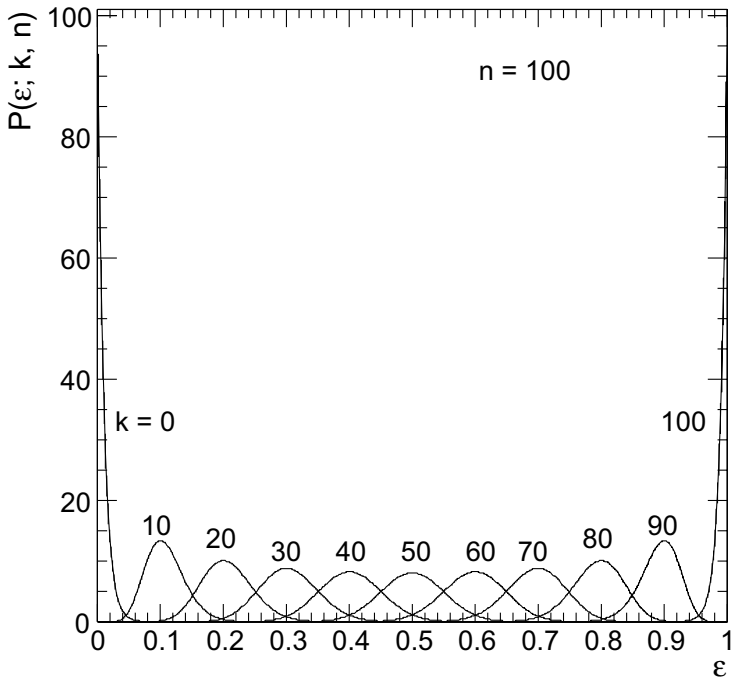


Figure 2: The probability density function $P(\varepsilon; k, n)$ for $n = 100$ and $k = 0, 10, \ldots, 100$.

course in the limit of large $n$. As depicted in Fig. 1 the density functions are skewed except for $k = n/2 = 5$.

What to use if only one measurement was made: mean or mode? Statistics teaches us that an estimator cannot be described as 'right' or 'wrong' but only as 'good' or 'bad'. A good estimator has to be consistent, unbiased, and efficient. Clearly $\hat{\varepsilon} = k/n$ has these features, while $(k+1)/(n+2)$ is biased for small n. In practice, however, this difference can be neglected since in most cases n is reasonably large. This is depicted in Fig. 2 for the case $n = 100$.

Let's now turn to the original purpose of this exercise: the error evaluation. The calculation of the variance $V(\varepsilon)$ from $P(\varepsilon; k, n)$ yields:

$$\begin{aligned} V(\varepsilon) &= \overline{\varepsilon^2} - \overline{\varepsilon}^2 && (17) \\ &= \int_0^1 \varepsilon^2 \, P(\varepsilon; k, n) \, d\varepsilon - \overline{\varepsilon}^2 && (18) \\ &= \frac{(k+1)\,(k+2)}{(n+2)\,(n+3)} - \frac{(k+1)^2}{(n+2)^2}. && (19) \end{aligned}$$

As expected $V(\varepsilon)$ now behaves correctly in the two extreme cases; for $k = 0$

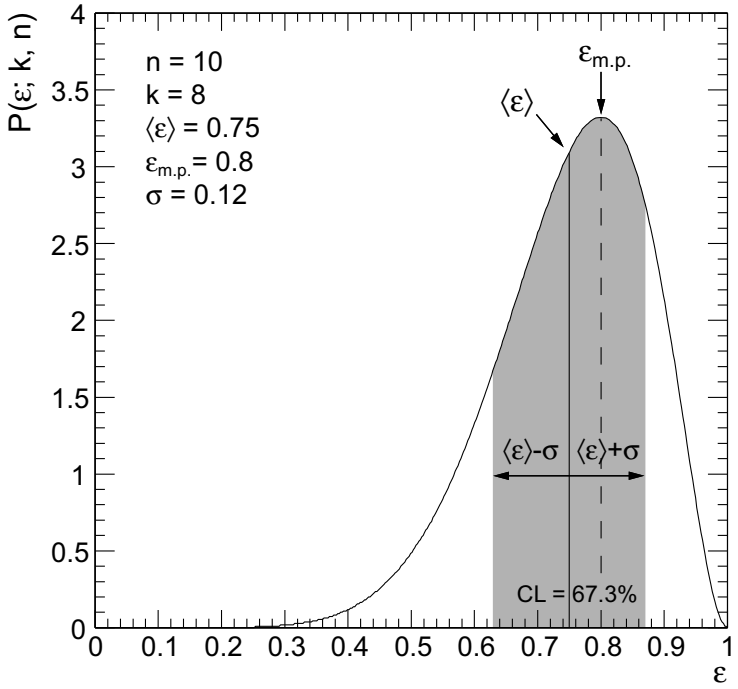


Figure 3: Efficiency probability density function $P(\varepsilon; 8, 10)$. The solid vertical line depicts the mean value, the dashed line the most probable value. The gray shaded region corresponds to plus/minus one standard deviation (Eq. 19) around the mean.

and $k = n$ one finds

$$V(\varepsilon)|_{k=0,n} = \frac{n+1}{(n+2)^2\ (n+3)}\ > 0. \tag{20}$$

For large $n$ in this case the variance becomes $\lim_{n\to\infty} V(\varepsilon) = 1/n^2$.

Interestingly the mean and variance for $n = 0$ is non-zero. One finds $\overline{\varepsilon} = 1/2$ and $V(\varepsilon)|_{n=0} = 1/12$, which is simply the mean and variance of a uniform distribution in the interval from 0 to 1. This is a nice confirmation of the validity of our calculations. The case $n = 0$ essentially means that no prior information on the efficiency is available and all we can say beforehand is that the efficiency has to be between 0 and 1. Since no efficiency is more likely than any other, the only assumption one can make is that of a uniform probability density.

Figure 3 shows the probability density function for the case $n = 10$ and $k = 8$. The vertical dashed and solid lines depict the mode and the mean, respectively. In this case the region $\overline{\varepsilon} \pm \sigma$ where $\sigma_\varepsilon = \sqrt{V(\varepsilon)}$ from Eq. 19, corresponds to a confidence level (CL) of 67.3%.

### 3.3 Measure for the statistical error

There are many measures of the uncertainty that can be extracted from the distribution $P(\varepsilon; k, n)$: 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. Whatever choice, the situation is complicated by the fact that the function is *(i)* not symmetric and *(ii)* that the integral over $P(\varepsilon; k, n)$ is not analytic for limits other than 0 and 1.

There is no obvious recipe for cases where $n$ is small and/or the efficiency found is close to 0 or 100%. A reasonable measure is certainly the standard deviation $\sigma_\varepsilon$. The advantage of this approach is clearly that $\sigma_\varepsilon = \sqrt{V(\varepsilon)}$ can be easily calculated from Eq. 19. In addition it turns out that the confidence levels for 'common' data sample sizes $n$ and efficiencies are close to the "$1\sigma$" probability content of the Gaussian distribution, *i.e.*, 68.3%. This is of course no accident but a consequence of the Central Limit Theorem. Note that in the limiting cases, that is, small $k$, small $n$, or $k$ close or equal to $n$, the confidence level deviates slightly from this value.

## 4 Acknowledgments

This report is the result of a discussion between the authors in the context of data analysis for the STAR experiment at the Relativistic Heavy-Ion Collider (RHIC) at BNL. After finalizing the calculations we discovered a writeup on this very subject by Marc Paterno from University of Rochester (D0 Note 2861). In this paper the author derives the same probability density function but does not calculate mean, mode, and standard deviation. He concludes that confidence levels cannot be found analytically and refers to a numerical solution in the form of a program.