On statistical errors of efficiencies and inefficiencies

Revised version

In our simulation or measurement we typically start N events. At the end these total number splits up in two distinct group of events, the detected ones (N_D) representing the efficiency of the detector (ϵ_D) and the missed one (N_M) representing the inefficiency (ϵ_M). How to calculate the statistical error of these values? If one now uses - straight forward - the standard rules on the formulas $\epsilon_D = N_D/N$ or $\epsilon_M = N_M/N$, respectively, this leads to wrongs and inconsistent number, because the sets N_D and N (and N_M and N, repectively) are not independent in term of set theory, i.e. they share a set of events, which are in both sets. More precise, N_D and N_M are subsets of N. Consequently, the error of N is not independent from the others.

The only thing one has to do is choosing the right variable, in this case the distinct sets N_D and N_M with independent errors. The rest is mathematics:

$$\epsilon_D = \frac{N_D}{N} = \frac{N_D}{N_D + N_M}$$

and

$$\epsilon_M = \frac{N_M}{N} = \frac{N_M}{N_D + N_M}.$$

This immediately fixes the problem, that the error bars could extend to below zero and above one, e.g.

$$\epsilon_D + \sigma_{\epsilon_D} = \frac{N_D + \sigma_{N_D}}{N_D + \sigma_{N_D} + N_M \pm \sigma_{N_M}} < 1,$$

whenever $N_M - \sigma_{N_M} \geq 0$ is ensured.

Following the error propagation (Gauß)

$$\sigma_g^2 = \sum_i \left(\frac{\delta g}{\delta x_i}\right)^2 \sigma_{x_i}^2$$

we get

$$\begin{aligned}
\sigma_{\epsilon_D}^2 &= \left(\frac{\delta \epsilon_D}{\delta N_D}\right)^2 \sigma_{N_D}^2 + \left(\frac{\delta \epsilon_D}{\delta N_M}\right)^2 \sigma_{N_M}^2 \\
&= \left(\frac{N_M}{(N_D + N_M)^2}\right)^2 \sigma_{N_D}^2 + \left(\frac{N_D}{(N_D + N_M)^2}\right)^2 \sigma_{N_M}^2 \\
&= \frac{1}{N^2}\left[(N_M \sigma_{N_D})^2 + (N_D \sigma_{N_M})^2\right]
\end{aligned}$$

For the standard deviation of ϵ_M you will get **exactly** the same like it should be.

If we are dealing with a considerable large number, a confidence level of $\approx 68\%$ is represented by one standard deviation $\sigma = \sqrt{N}$. Therefore

$$\sigma_{\epsilon_D} = \frac{1}{N^2}\sqrt{N_M^2 N_D + N_D^2 N_M}.$$

If we have $N_D = 1000$ and $N_M = 5000$, then we will get:

$$\begin{aligned} \epsilon_D &= (16.7 \pm 0.5)\% \\ \epsilon_M &= (83.3 \pm 0.5)\% \end{aligned}$$

Another complication comes into the game if we are approaching $N_D = 0$ or $N_M = 0$. Here, $\sqrt{N}$ does no longer coincide with a confidence level of 68% (see Rev. of Part. Physics, sect. 28.6.4.)[1]. In practice, this only gets important for $N = 0$. Here, a confidence level of 68% means $\Delta N = 1$. For example, having $N = N_D$ and $N_M = 0$, the equation above can still be used with $\sigma_{N_M} = 1$ and taking the result as an lower (upper) limit:

$$\begin{aligned} \sigma_{\epsilon_D} &= \frac{1}{N^2}\sqrt{N_M^2\sigma_{N_D}^2 + N_D^2\sigma_{N_M}^2} \\ &= \frac{1}{N^2}\sqrt{0 \cdot \sqrt{N} + N^2 \cdot 1} \\ &= \frac{1}{N} \end{aligned}$$

Thus, $\sigma_{\epsilon_D} \neq \frac{1}{\sqrt{N}}$ as one might think of as a first guess. For $N = 100$, e.g., we have then:

$$\begin{aligned} \epsilon_D &> 99\% \quad , \quad \mathrm{CL} = 68\% \\ \epsilon_M &< 1\% \quad , \quad \mathrm{CL} = 68\% \end{aligned}$$

V. Hejny, Nov. 2001

[1] To be more precise, one has to use a Poisson distribution instead of a Gaussian one for small samples. Then it gets also obvious, that with small numbers, peaks are always more pronounced than dips.