Efficiency measurement: a Bayesian approach

Diego Casadei[a]

[a] *Physics Department, New York University,*
4 Washington Place, New York, NY-10003, USA

Abstract

The measurement of the efficiency of some event selection is always an important part of the analysis of experimental data. The statistical techniques based on the use of the Bayes theorem which are needed to determine the efficiency and its uncertainty are reviewed. The problem of choosing a meaningful prior is addressed, and different priors are considered in real-life use cases. The use of the uncertainties in practical cases is also considered, together with the problem of combining different samples. "Pathological" cases are also addressed, in which non-unit weights or non-independent selections have been used to fill the histograms. The use of the family of Beta distributions is illustrated in the examples, showing how its conjungate property for binomial sampling makes it the most convenient choice for defining priors. Finally, several recommendations are made about the choice of the prior and about using and communicating the results.

Key words: efficiency, Bayesian approach, reference analysis, non uniform priors
PACS: 02.70.Rr, 06.20.Dk, 07.05.Kf, 29.85.Fj

1. Introduction

There are several cases in high-energy physics in which we are interested into measuring an efficiency, for example when dealing with the trigger or offline event selection with the aim of measuring a cross section. With our selection, we reject a subset of the input data set and we look at the ratio between the number of surviving events and the initial number. The selection efficiency is interpreted as the probability that any single event passes the selection. This statement is independent from the actual definition of

Email address: `diego.casadei@cern.ch` (Diego Casadei)

probability (provided that it satisfies all required properties), but in this paper we interpret it as the degree of belief that some statement is true. This interpretation is quite natural and is needed for the Bayes' theorem to be able to tell us about the probability distribution of the true but unknown efficiency, which we want to estimate.

When speaking about a selection process, we count the initial number of events and the final number of selected events. Usually, one also knows the distribution (actually, the histogram) of some control parameters before and after the selection, and wants to determine the efficiency as function of such parameters. In this paper, the aim is to show how to estimate the efficiency (possibly as function of the parameters) and its uncertainty within the framework of Bayesian statistics. The Bayesian approach is also illustrated with several examples which can serve as guidance for most problems daily encountered by the experimenters.

The frequentist approach, often preferred in high-energy physics data analysis, is not addressed here — the reader is kindly invited to find the recent review by [1] and the short comparison between the two approaches in [2, chap. 32] — for the following reasons.

The experimenters most often ask about the probability that, given the data, the true value assumes some value. This question is answered by the Bayesian approach but is ill-defined in the frequentist approach, because the true quantity (the efficiency) is viewed as a fixed unknown parameter which does not "fluctuate", so that a probability distribution cannot be defined. On the other hand, in the Bayesian framework the probability statements always refer to some state of knowledge, so that it is possible (and indeed it is desired) to define a probability distribution for the true quantity, interpreted as a description of our degree of belief about its unknown value.

The frequentist solution is often expressed as a "confidence interval" with some degree of "coverage". The two concepts are closely related to the ideal repetition of the same experiment, exactly in the same conditions, for a large number of times: for example, a 95% confidence interval is expected to contain the true (unknown) value of the parameter for (not less than) 95% of the repeated experiments, while (at most) 5% of them will miss it. However, in practice repeated experiments with identical conditions are only possible in computer simulations. In real cases, this long sequence of measurements is either (1) impossible — because the experiment is too complex and/or too expensive, or because perfect stability in time is impossible to achieve — or (2) differently interpreted as a single and longer experiment — this is done

every day in high-energy physics, when people collect together several runs to get higher statistics rather than interpreting them as repeated measurements —. Hence, in this paper the interval coverage is not considered a fundamental property, though it will be mentioned again later on (in section 4).

In the Bayesian approach, as well as in the frequentist one, the likelihood (i.e. the probability that an event occurs, given the model) plays a central rôle. However, expecially when the input number of events is very low, one must also pay attention to the "prior" probability density function (PDF), which represents our degree of belief about the possible values of the unknown efficiency *before* the experiment is actually carried out. Indeed, the full Bayesian solution is provided by the "posterior" PDF, which is proportional to the product of the likelihood with the prior (Bayes' theorem). Unless the prior is pathologic (i.e. null or negligible in the region where the true value is), the likelihood will "attract" the posterior more and more, as the number of input events increases.

When the full posterior PDF cannot be used, it can be summarized by providing its mean and some measure of its dispersion (usually the standard deviation), which can be used in the computation of the desired physical quantities in the usual way. The only caveat is that working with the standard deviation implies assuming that the underlying PDF is somewhat symmetric about the mean, which in some case might be a bad approximation (see section 2.2). In general, it is recommended to work with the full posterior whenever it is possible.

Note that a confidence interval found with a frequentist approach cannot be used together with the best estimate (usually, the maximum-likelihood estimate or MLE) as a "two-sigma" range, because it cannot be interpreted as the width of a probability distribution[1]. Hence, when the efficiency is needed as intermediate step in the computation of some physical quantity, the frequentist solution cannot be used to compute the uncertainty on such quantity with the usual "error propagation". In contrast, the Bayesian posterior mean and variance can be used to compute the final uncertainty in the usual way, and this is a further reason to consider the Bayesian approach.

In the rest of the paper, the problem of estimating the selection efficiency and its uncertainty (a concept which only makes sense in the Bayesian frame-

[1]This is often a source of confusion among experimenters, used to play with normal distributions and to interpret the results in a (unconscious) Bayesian way.

work) is addressed. The choice of the prior requires some care, so that recommendations are made in section 3 to deal with the usual practical cases of informative and non-informative priors. Few cases in which the assumptions of independent selections with unit weight are not both satisfied are addressed in section 5. In addition, the problem of fitting parametric models is addressed in section 6. Finally, useful mathematical relations and more complicate developments are shown in the appendix.

2. How to measure the efficiency and its uncertainty?

Being a probability, the efficiency cannot be directly measured. Instead, we must estimate it with the available data: we can only count events, i.e. measure *relative frequencies*, that are rational numbers. By virtue of the weak law of large numbers (or Bernoulli's theorem), the relative frequency will converge in probability to the true efficiency in the limit of an infinite number of measurements. Convergence in probability implies that "unusual" outcomes become less and less likely as the sequence $\{X_n\}$ of random variables progresses, and is formally expressed by saying that, for any small positive number $\epsilon > 0$, the probability that the distance between X_n and the true value X exceeds ϵ tends to zero: $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$.

It is important to emphasize that the tendency of the measured frequency to the probability has *not* the same behavior of a mathematical limit, for which it never happens, after some point, that the distance to the limit exceeds a given small quantity: convergence in probability is weaker. Indeed, it is always possible to find, even after a large number of trials, a frequency that is not very near to the probability, although the *probability* for this to happen decreases with increasing number of trials (Bernoulli's theorem). This makes measuring probabilities conceptually different from measuring physical quantities like e.g. the electric field in some point, that is given by the mathematical limit of the measured force divided by the test charge, when the latter goes to zero. However, in practice this difference is not very important because of the fluctuations of stochastic nature which affect the measurement process, and this might explain why many good books on statistical methods of data analysis (e.g. [3]) do not make this distinction.

Usually, we create and fill histograms, assuming that they converge to the true distributions in the limit of infinite statistics and zero bin width

as the partial sums converge to the Riemann's integral.[2] Hence, we usually approximate the PDF $\varepsilon(x; A)$ describing the efficiency of the selection A as function of the parameter(s) x, with the step function representing the observed relative frequencies $\{f_i(A)\}$:

$$\varepsilon_i \equiv \int_{x_i}^{x_{i+1}} \varepsilon(x; A)\, \mathrm{d}x \approx f_i(A) \equiv \frac{k_i(A)}{n_i} \ .$$

where n_i and $k_i(A)$ count the entries in the i-th bin before and after the selection A.

2.1. The binomial distribution

A histogram of a quantity x obtained with a series of repeated measurements of x is a collection of pairs (i, n_i) representing the number n_i of times the measured value of x has been found in the i-th bin. The integer values n_i have been observed, hence they have no uncertainty. However, if we consider the histogram as an estimate of the true distribution of x, then the n_i's are estimates of the integral of the true distribution in each bin. In the assumption that the populations of all bins are statistically independent, the uncertainty σ_i of the estimate n_i for the true population n_i^* of bin i is given by the Poisson distribution: $\sigma_i = \sqrt{n_i}$.

This assumption is usually justified, but in our case the entries n_i and k_i of bin i before and after the selection are *not* statistically independent, hence one can not compute the variance of f_i with the usual rules of the "propagation of errors".[3] Rather, the application of a selection on each bin can be considered a binomial process, with probability of "success" ε_i, the true (but unknown) efficiency: the probability to obtain k_i events passing the selection when the efficiency is ε_i and the sample size is n_i is:

$$P(k_i|\varepsilon_i, n_i) = \binom{n_i}{k_i} (\varepsilon_i)^{k_i} (1 - \varepsilon_i)^{n_i - k_i} \equiv \mathrm{Bi}(k_i|\varepsilon_i, n_i) \tag{1}$$

with mean $E(k_i|\varepsilon_i, n_i) = \varepsilon_i n_i$ and variance $V(k_i|\varepsilon_i, n_i) = \varepsilon_i(1 - \varepsilon_i)n_i$. However, this does not solve our problem, because ε_i is still unknown (instead, we measured n_i and k_i).

[2] Again, this is not rigorously correct: the histograms converge only in probability.

[3] Please, note that this is the default behavior of the bin-wise histogram division performed by ROOT [4], unless special options are provided.

2.2. The usual (abused) approximation

In practice, often people compute $V(k|n,\varepsilon) = n\varepsilon(1-\varepsilon)$ from (1) — we drop the index i for a while — and find the approximate variance for the relative frequency $f = k/n$ by dividing by n^2, which has no uncertainty because it has been measured. Finally, the result is approximated[4] by the substitution $\varepsilon \to f$:

$$V(f) = \frac{V(k)}{n^2} = \frac{\varepsilon(1-\varepsilon)}{n} \approx \frac{f(1-f)}{n} = \frac{k(n-k)}{n^3} \,. \tag{2}$$

Though this is a good approximation when both n, k are large and the observed frequency is not too similar to zero or one, rigorously speaking the formula (2) is incorrect because the binomial distribution (1) is a function of k_i with parameters ε_i and n_i whereas in our case n_i and k_i are both known and we want to find ε_i: we should look instead for a function of ε_i, with parameters n_i and k_i, as it will be shown in section 2.3. In addition, this approximation suffers from the following problem: the formula (2) gives zero uncertainty for the two limiting cases $k = 0$ and $k = n$, independently from the actual value of n. This means that, if we have a single event ($n = 1$) and this survives the cut, we get the very same result (zero uncertainty) as the case $k = n = 100$, whereas one would expect the latter estimate to be (roughly 10 times) more precise.

2.3. The probability distribution for the true efficiency

In order to find a function of ε_i, given n_i and k_i, we use the Bayes' theorem: $P(\varepsilon_i|k_i,n_i) \propto P(k_i|\varepsilon_i,n_i)P(\varepsilon_i|n_i)$. We know k_i and n_i and that the process is binomial[5], so that the likelihood is $P(k_i|\varepsilon_i,n_i) = \mathrm{Bi}(k_i|\varepsilon_i,n_i)$ from equation (1). In addition, the prior should not depend on the sample size, $P(\varepsilon_i|n_i) = P(\varepsilon_i)$, hence

$$P(\varepsilon_i|k_i,n_i) = \frac{1}{C_i}\,\mathrm{Bi}(k_i|\varepsilon_i,n_i)\,P(\varepsilon_i) \tag{3}$$

(where C_i is a normalization constant) is the probability that the true efficiency in bin i is between ε_i and $\varepsilon_i + \mathrm{d}\varepsilon_i$.

[4]This approximation is used by ROOT when the option `"B"` is given to the histogram division.

[5]The histograms *must* be filled with unit weight for this to be true.

3. The choice of the prior

The prior encodes our state of knowledge before the measurement is carried out. Often, either we precisely know the prior PDF or we want to model a state of "perfect ignorance". The first case happens for example when we already know the efficiency of some device from a recent calibration procedure, while the second case could be the attempt of providing the most "objective" estimate of the efficiency. Intermediate cases in which we have some vague prior expectation in mind are also possible.

3.1. The Beta distribution

If we have some knowledge before we measure the relative frequency, it should be encoded into a specific form for the prior PDF. One does not need to be very precise, because the Bayes' theorem assures that the final result (the posterior probability) will be driven by the data, provided that we have enough events and that the prior is not null or negligible in the region containing the true value.

In our case, the recommended family for the prior PDF is the Beta family, whose properties are listed in appendix A.2, because its natural conjungate property for binomial sampling of k successes among n trials brings a Beta prior with parameters a, b into a Beta posterior with parameters $a' = a + k$ and $b' = b + n - k$. Simple formulae exist to express the mean, mode and variance of the resulting posterior distribution, as shown in appendix A.2, so that we can use the "method of moments" to find the Beta parameters which best match our prior knowledge of the value and uncertainty of the true efficiency. We can assume that such values are the mean E and variance V of the prior PDF and find the Beta parameters a, b as

$$a = E\left[\frac{E(1-E)}{V} - 1\right] \tag{4}$$

$$b = (1-E)\left[\frac{E(1-E)}{V} - 1\right] \tag{5}$$

to model the prior as a Beta distribution. Examples of the use of the method of moments to define an approximate prior PDF are shown in the following sections.

3.2. Non-informative priors

The choice of a uniform prior (a Beta distribution with $a = b = 1$) could seem "non-informative" (or better the least informative one) and appropriate if we have no prior knowledge of $P(\varepsilon)$ and we do not have any reason to prefer any particular range for the true efficiency [5, 6]. However, one may also think in terms of the logarithm (or some other function) of the efficiency and want a non-informative prior also for such parametrization. Unfortunately, the uniform prior is not invariant under reparametrization, so that the prior written as function of $\log \varepsilon$, for example, is no more uniform (it is multiplied by the Jacobian determinant). In summary, the uniform prior is a legitimate *informative* prior when we have good reasons to use it, but is not a correct choice when one wants to model the situation of minimal prior information.

The goal of the Bayesian "reference analysis" [7] is to study the impact of the choice of the prior on the posterior PDF, with respect to the case of the minimal possible prior knowledge. The *reference posterior* is the Bayesian result which minimally depends on the prior knowledge (equivalently, it maximally depends on the likelihood), and the corresponding *reference prior* [8] is used in the Bayes' theorem to model the minimal information on the system before carrying out the experiment. One may view the reference posterior as the "most objective" Bayesian solution, and consider using it whenever a result should be published in a way which minimally depends on the prior experimenter knowledge, or when it is desired to assess the dependency of the solution on the choice of the prior — the reader may find all details about the use of reference analysis as a tool for objective Bayesian inference in [9].

There are good reasons to choose the least informative prior. Often the minimal prior information is indeed the best model, but when this is not the case one certainly wants to assess the dependence of the solution (the posterior PDF) on the choice of the prior. One of the requirements for the reference analysis is the invariance under reparametrization. For the binomial case, it comes out that the reference prior coincides with the Jeffreys' prior (a Beta distribution with $a = b = 1/2$, proportional to $\varepsilon^{-1/2}(1-\varepsilon)^{-1/2}$), which was indeed found by requiring invariance under reparameterization [10]. The Jeffreys' prior is the recommended choice to model the least informative prior in our efficiency study, and the resulting posterior PDF, $\mathrm{Be}(\varepsilon; k + 0.5, n - k + 0.5)$, is the reference posterior for the binomial case. Figure 1 shows the reference posterior for a sample size of $n = 10$ and $n = 100$.

From the relations (31) of appendix A.2 we immediately get the following values for the mean, mode, variance and skewness of the reference posterior

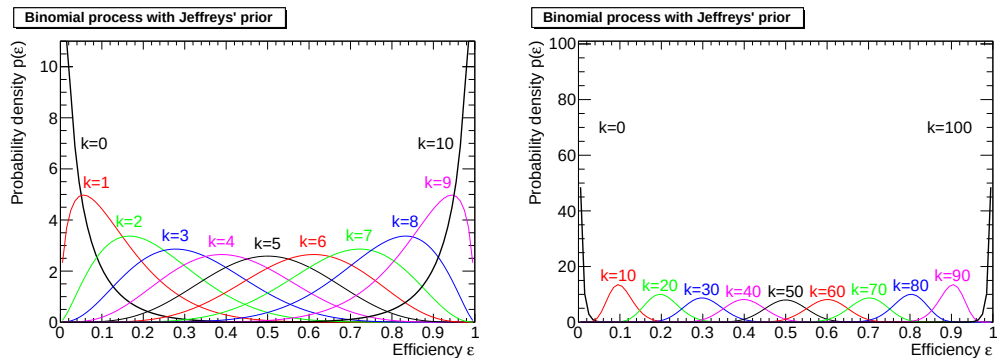

Figure 1: Reference posterior probability density function $\mathrm{Be}(\varepsilon; k+0.5, n-k+0.5)$ for $n = 10$ (left) and $n = 100$ (right).

$\mathrm{Be}(\varepsilon; k+0.5, n-k+0.5)$:

$$E(\varepsilon) = \frac{k+0.5}{n+1} \tag{6}$$

$$m(\varepsilon) = \frac{k-0.5}{n-1} \tag{7}$$

$$V(\varepsilon) = \frac{(k+0.5)(n-k+0.5)}{(n+1)^2\,(n+2)} \tag{8}$$

$$\gamma_1(\varepsilon) = \frac{2(n-2k)\sqrt{n+2}}{(n+3)\sqrt{(k+0.5)(n-k+0.5)}} \tag{9}$$

(the mode is only defined for $n > 1$, which is also an obvious requirement for the actual measurement). The result is that both the mean and the mode are biased (but robust) estimators of the efficiency. In general, the distribution is asymmetric ($\gamma_1 \neq 0$) and the mode is different from the mean, apart from the case $n = 2k$. Note that for $k = 0, n$ the variance is not null

$$V(\varepsilon)|_{k=0,n} = \frac{0.5(n+0.5)}{(n+1)^2(n+2)} > 0$$

and $V(\varepsilon)|_{k=0,n} \propto 1/n^2$ when $n \to \infty$, as expected.

3.3. Jeffreys' versus uniform prior

Both Paterno [5] and Ullrich & Xu [6] considered as a reasonable choice for the prior $P(\varepsilon_i)$ a uniform distribution in $[0, 1]$, though this is questionable

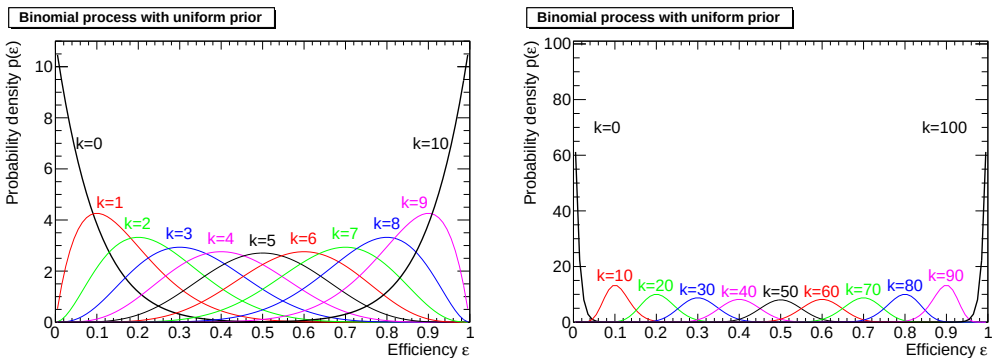

Figure 2: Posterior PDF $P(\varepsilon|k,n)$ obtained with the uniform prior, for $n = 10$ (left) and $n = 100$ (right).

because it does *not* represent a "complete ignorance", as mentioned above (section 3.2). On the other hand, the choice of the uniform prior is interesting at least for two reasons. First, the posterior PDF $\mathrm{Be}(\varepsilon; k+1, n-k+1)$, shown in figure 2 with 10 and 100 initial events, is proportional to the likelihood, so that the MLE (i.e. the frequentist solution) coincides with the mode of the posterior PDF. Second, credible intervals obtained with such posterior have been implemented in ROOT, being the only available Bayesian credible regions in such framework.

The mean of $\mathrm{Be}(\varepsilon; k+1, n-k+1)$ is

$$E(\varepsilon) = \int_0^1 \varepsilon\, P(\varepsilon|n,k)\, \mathrm{d}\varepsilon = \frac{k+1}{n+2} \tag{10}$$

and is a bit more biased estimator than the reference mean (10). The mode, i.e. the value at which $\mathrm{d}P/\,\mathrm{d}\varepsilon = 0$, is

$$\mathrm{mode}(\varepsilon) = \frac{k}{n}\ .$$

and indeed it coincides with the MLE $f = k/n$. Again, because of the shape asymmetry, this does not coincide with the expectation value, apart for the case $f = 0.5$. Finally, the variance is:

$$V(\varepsilon) = \frac{(k+1)(n-k+1)}{(n+2)^2\,(n+3)}\ . \tag{11}$$

Incidentally, figure 2 is also useful to check by eye if the symmetric binomial approximation of section 2.2 is well justified, because the only difference

between the posterior $\mathrm{Be}(\varepsilon; k+1, n-k+1)$ and the binomial likelihood is just the normalization (the likelihood is not normalized),

The reference posterior (obtained using the Jeffreys' prior) is not much different from the posterior obtained with the uniform prior, unless n is relatively small. The difference with respect to the frequentist MLE is usually larger, and is most apparent when the measured relative frequency is one or zero (which are never the best estimate in the Bayesian approach), though is less significant for the reference posterior.

Table 1 shows the mean (or MLE) and square root of the variance for all cases, together with the ratio between the posterior mean and variance obtained with the uniform prior and the corresponding quantity computed with the reference posterior. Clearly, the biggest discrepancy between the results obtained with the uniform and Jeffreys' priors is obtained when $k = 0$ and increases for higher n, because the expected value using the uniform prior is $(n+2)^{-1}$ and the reference posterior mean is $(2n+2)^{-1}$. The ratio between the standard deviations is larger than one at very small and large k values and smaller when the efficiency is intermediate, whereas the uniform mean is larger than the reference posterior mean for small relative frequencies and smaller for large relative frequencies (the means are equal only when $n = 2k$). Given that the reference posterior is "more objective" and that its mean is a less biased estimator of the true efficiency, the recommendation is to use the Jeffreys' prior unless there are good reasons to use the (informative) uniform prior.

3.4. Combining independent samples

If we have two independent efficiency measurements for the same process and we want to use all available information, the correct approach to combine them is to merge the samples before and after the selection and use the results to make the final estimate (this is also true for the frequentist approach). In the Bayesian approach, the very same result is also obtained if we use the first posterior PDF to model the prior for the second estimate. Indeed, the Bayes' theorem can be interpreted as a model for our learning process: it makes use of all information available at any given time, a very desirable property.

To make an example, let us consider the histograms $\{(i, k_i)\}$ and $\{(i, k'_i)\}$, filled with all events in the first sample and with the subset obtained after the selection, and the analogous histograms $\{(i, n_i)\}$ and $\{(i, n'_i)\}$ filled before and after the cut with the second sample. Here we assume that all histograms

		Binomial app.		**Uniform prior**		**Jeffreys' prior**		**Unif./Jeffreys**	
n	k	**MLE**	**s. d.**	**mean**	**s. d.**	**mean**	**s. d.**	**mean**	**s. d.**
2	0	0.000	0.000	0.250	0.194	0.167	0.186	1.500	1.039
2	1	0.500	0.354	0.500	0.224	0.500	0.250	1.000	0.894
2	2	1.000	0.000	0.750	0.194	0.833	0.186	0.900	1.039
3	0	0.000	0.000	0.200	0.163	0.125	0.148	1.600	1.104
3	1	0.333	0.272	0.400	0.200	0.375	0.217	1.067	0.924
3	2	0.667	0.272	0.600	0.200	0.625	0.217	0.960	0.924
3	3	1.000	0.000	0.800	0.163	0.875	0.148	0.914	1.104
4	0	0.000	0.000	0.167	0.141	0.100	0.122	1.667	1.150
4	1	0.250	0.217	0.333	0.178	0.300	0.187	1.111	0.952
4	2	0.500	0.250	0.500	0.189	0.500	0.204	1.000	0.926
4	3	0.750	0.217	0.667	0.178	0.700	0.187	0.952	0.952
4	4	1.000	0.000	0.833	0.141	0.900	0.122	0.926	1.150
5	0	0.000	0.000	0.143	0.124	0.083	0.104	1.714	1.184
5	1	0.200	0.179	0.286	0.160	0.250	0.164	1.143	0.976
5	2	0.400	0.219	0.429	0.175	0.417	0.186	1.029	0.939
5	3	0.600	0.219	0.571	0.175	0.583	0.186	0.980	0.939
5	4	0.800	0.179	0.714	0.160	0.750	0.164	0.952	0.976
5	5	1.000	0.000	0.857	0.124	0.917	0.104	0.935	1.184
10	0	0.000	0.000	0.083	0.077	0.045	0.060	1.833	1.275
10	1	0.100	0.095	0.167	0.103	0.136	0.099	1.222	1.043
10	2	0.200	0.126	0.250	0.120	0.227	0.121	1.100	0.993
10	3	0.300	0.145	0.333	0.131	0.318	0.134	1.048	0.972
10	4	0.400	0.155	0.417	0.137	0.409	0.142	1.019	0.963
10	5	0.500	0.158	0.500	0.139	0.500	0.144	1.000	0.961
10	6	0.600	0.155	0.583	0.137	0.591	0.142	0.987	0.963
10	7	0.700	0.145	0.667	0.131	0.682	0.134	0.978	0.972
10	8	0.800	0.126	0.750	0.120	0.773	0.121	0.971	0.993
10	9	0.900	0.095	0.833	0.103	0.864	0.099	0.965	1.043
10	10	1.000	0.000	0.917	0.077	0.955	0.060	0.960	1.275

Table 1: Comparison between the posterior means and standard deviations obtained with the uniform and Jeffreys' priors. The frequentistic MLE and the standard deviation in the binomial approximation are also shown.

are accessible, however the result is also valid if one of them is missing and we are given the posterior PDF which summarizes the unavailable measurement.

We take a single bin (omitting its index) and assume that the prior PDF for the first measurement is a Beta density with parameters a, b (in case of no prior knowledge we recommend using the Jeffreys' prior with $a = b = 0.5$), so that the posterior of the first measurement is the Beta density $\mathrm{Be}(\varepsilon; a+k', b+k-k')$. This then is used as the prior for the second estimate,

whose posterior PDF is the Beta density $\mathrm{Be}(\varepsilon; a+k'+n', b+k-k'+n-n')$. The very same posterior is obtained if we start from the initial prior $\mathrm{Be}(\varepsilon; a, b)$ and consider the joint sample of size $m = k+n$ from which only $m' = k'+n'$ events survived.

With the same method, we can also use simulated data to provide the prior distribution for the true efficiency, to be used in conjunction with real data in the Bayesian approach. This is mostly useful in case we need to test different kinds of "systematic" effects on the outcome of a real experiment.

Given that using densities belonging to the Beta family allows to summarize all available information in a rather easy way, the recommended way of communicating efficiencies is to provide the values of the 4 parameters of the Beta posterior and of the Beta prior. This is equivalent to the knowledge of the original samples (before and after the selection) and can be used to make a combined estimate of the efficiency without the original data: it is sufficient to use the corresponding Beta distribution as prior for the new measurement. In addition, the users will be able to test the sensitivity of the result to the choice of a different prior.

4. Quantifying the uncertainty

When the efficiency is needed to convert the measured quantities into the true ones (to get e.g. the cross section), its best estimate is needed together with its uncertainty (whenever using the full distribution is not practical). The mean and variance of the posterior PDF can be used in computing quantities following the usual recipe of the "error propagation". The only caveat is that, in general, the posterior is asymmetric so that the trivial recipe of taking "$\pm n\sigma$" intervals around the expected result might produce intervals which extend into an unphysical region. The asymmetry is more pronounced when the efficiency is very near to one or zero, and is negligible when the binomial approximation of section 2.2 is good (i.e. when both $k, n \gg 1$). In general, it is recommended to work with the full posterior whenever it is possible, expecially when the (symmetric) binomial approximation is poor.

Because physicists are used to take the interval $f_i \pm \sigma_i$, with $\sigma_i = \sqrt{V(\varepsilon_i)}$ as the Gaussian credible interval with 68.3% probability, this feature is often desired: Paterno [5] recommends using the smallest interval $[a, b] \subset [0, 1]$ that contains the probability $\lambda = 68.3\%$, i.e. the shortest credible interval with posterior probability λ (known in the statistics literature as the "highest posterior density" or HPD region), arguing that in practical applications

it will behave more or less as the "$\pm 1\sigma$" interval defined with a Gaussian standard deviation. He also provided code to compute such interval when using the uniform prior, which has been adopted by ROOT.

We emphasize that this prescription is not invariant under reparametrization (such credible interval is no more the HPD region after a general change of variable), which is a very desirable feature. It is recommended instead to show an invariante region like the 95% "reference credible intervals" [11] in the efficiency graph (see section 4.1 below).

Showing asymmetric errors is the recommended style for efficiency graphs, because the posterior PDF in general is not symmetric. The ROOT method TGraphAsymmErrors::BayesDivide() can be used to plot efficiency graphs showing Paterno's intervals (obtained with the uniform prior) as asymmetric errors on the relative frequencies. Unfortunately, it does not support non-integer input (ROOT 5.26/00 is considered here), so that it cannot be used to plot intervals corresponding to any possible choice for the prior[6] (most notably, it cannot be used with the Jeffreys' prior). In addition, Paterno's intervals are not invariant under reparametrization. So far, there is no available ROOT method for plotting reference credible histograms, though RooStats::BayesianCalculator could be a starting point to find central credible intervals. The latter contain the same probability on the left and on the right of the expected value and are invariant.

4.1. Invariant credible regions

In this section we summarize the treatment by Bernardo [11]. A possible choice for a credible interval is the *lowest posterior loss* or LPL credible region, which depends on the definition of the *loss function* which specifies the loss to be suffered if a particular value for the efficiency is used in place of the true value. To obtain a LPL credible region which is invariant under reparametrization, the loss function should depend on the full PDF, not on the value of a parameter. For example, the common choice of the quadratic distance $\ell = (\varepsilon_0 - \varepsilon)^2$ does not lead to an invariant solution.

Let us consider a model with probability distribution $p(x|\theta)$ for the observations $x \in \mathcal{X} \subseteq \mathbb{R}^n$, $n \geq 1$, dependent on the parameters $\theta \in \Theta \subseteq \mathbb{R}^m$, $1 \leq m \leq n$. An *intrinsic loss function* is a symmetric non-negative function

[6] If it happens that the Beta parameters a, b are integers, one can use TGraphAsymmErrors::BayesDivide() to plot 68.3% credible regions by passing it modified histograms filled with $k_i = a_i - 1$ and $n_i = b_i + a_i + 2$.

$\ell(\theta_0, \theta)$ which is zero if and only if $p(x|\theta_0) = p(x|\theta)$ almost everywhere in $\mathcal{X}$. An intrinsic loss function is invariant under reparametrization but not, in general, under a one-to-one transformation of the observations x. Because this is a very desirable property, we restrict ourselves to the intrinsic loss functions which are also invariant under one-to-one transformations of x. An example from this class is the the $\mathcal{L}_1$ norm, that is the integral of the absolute value of the difference between two distributions, computed at the same point x, over the whole support $\mathcal{X}$. When applied to the reference posterior for the binomial case, the $\mathcal{L}_1$ norm gives the invariant expected loss

$$\ell_1\{\varepsilon_0|k,n\} = \int_0^1 \ell_1(\varepsilon_0,\varepsilon)\,\mathrm{Be}(\varepsilon|k+\tfrac{1}{2}, n-k+\tfrac{1}{2})\,\mathrm{d}\varepsilon \tag{12}$$

$$\ell_1(\varepsilon_0,\varepsilon) = \sum_{k=0}^{n} |\,\mathrm{Bi}(k|\varepsilon_0,n) - \mathrm{Bi}(k|\varepsilon,n)\,| \tag{13}$$

independent from one-to-one transformations of ε. One can now build a LPL q-credible region by finding the interval $[\varepsilon_{\mathrm{low}}, \varepsilon_{\mathrm{up}}] \subset [0,1]$ which minimizes (12) under the constraint $\int_{\varepsilon_{\mathrm{low}}}^{\varepsilon_{\mathrm{up}}} \mathrm{Be}(\varepsilon|k+\frac{1}{2}, n-k+\frac{1}{2})\,\mathrm{d}\varepsilon = q$.

The behaviour of many important limiting processes in probability theory and statistical inference is better described in terms of another measure of divergence, related to the information theory, the *intrinsic discrepancy* $\delta\{p_1,p_2\}$, defined in terms of the Kullback-Leibler *directed divergence* $\kappa\{p_1,p_2\}$ between two PDFs p_1, p_2:

$$\delta\{p_1,p_2\} = \min\{\kappa\{p_1,p_2\}, \kappa\{p_2,p_1\}\} \tag{14}$$

$$\kappa\{p_i,p_j\} = \int_{\mathcal{D}} p_i(x)\log\frac{p_i(x)}{p_j(x)}\,\mathrm{d}x \tag{15}$$

The intrinsic discrepancy is symmetric, non-negative, defined for strictly nested supports, invariant under one-to-one transformations, and additive for independent observations. It may be viewed as the minimum expected log-likelihood ratio in favour of the model which generates the data (the "true" model, which is assumed to be described either by p_1 or p_2) and can be used to defined the *intrinsic discrepancy loss*

$$\delta_x\{\theta_0,\theta\} = \delta\{p(x|\theta_0), p(x|\theta)\} \tag{16}$$

where θ is the parameter in which we are interested. For the binomial model,

the intrinsic discrepancy loss is

$$\delta_k\{\varepsilon_0, \varepsilon|n\} = n\,\delta_x\{\varepsilon_0, \varepsilon\} \tag{17}$$

$$\delta_x\{\varepsilon_0, \varepsilon\} = \min\{\kappa\{\varepsilon_0|\varepsilon\}, \kappa\{\varepsilon|\varepsilon_0\}\} \tag{18}$$

$$\kappa\{\varepsilon_i|\varepsilon_j\} = \varepsilon_j \log\frac{\varepsilon_j}{\varepsilon_i} + (1-\varepsilon_j)\log\frac{1-\varepsilon_j}{1-\varepsilon_i} \tag{19}$$

where $\delta_x\{\varepsilon_0, \varepsilon\}$ is the intrinsic discrepancy between Bernoulli random variables with parameters ε_0 and ε.

Finally, *intrinsic credible regions* are defined as the lowest posterior loss credible regions which correspond to the use of the intrinsic discrepancy loss function together with the reference posterior. The reference posterior expected loss from using ε_0 rather than ε in the binomial model is

$$d\{\theta_0|k, n\} = n\int_0^1 \delta_x\{\varepsilon_0, \varepsilon\}\mathrm{Be}(\varepsilon|k+\tfrac{1}{2}, n-k+\tfrac{1}{2})\,\mathrm{d}\varepsilon \tag{20}$$

and the intrinsic q-credible interval is the interval $[\varepsilon_{\mathrm{low}}, \varepsilon_{\mathrm{up}}] \subset [0,1]$ which minimizes the loss (20) under the constraint $\int_{\varepsilon_{\mathrm{low}}}^{\varepsilon_{\mathrm{up}}} \mathrm{Be}(\varepsilon|k+\frac{1}{2}, n-k+\frac{1}{2})\,\mathrm{d}\varepsilon = q$.

Approximate expressions for the intrinsic credible intervals are based on the asymptotic expressions obtained in theorem 4.1 of Bernardo [11]. In particular, they are built upon the *reference parametrization* $\phi = \phi(\theta)$ of the parameters θ of interest, defined as the one for which the reference prior is uniform. For the binomial model there is a single parameter ε and $\phi(\varepsilon) = 2\arcsin\sqrt{\varepsilon}$, i.e. $\varepsilon(\phi) = \sin^2(\phi/2)$. Using a shorter notation for the reference posterior mean $\mu = E(\varepsilon) = (k+0.5)/(n+1)$ and variance $\sigma^2 = V(\varepsilon) = \mu(1-\mu)/(n+2)$ of the parameter of interest, the variance of the reference parametrization is

$$\sigma_\phi^2 \approx \sigma^2[\phi'(\mu)]^2 = \frac{1}{n+2} \tag{21}$$

while its mean is

$$\begin{aligned}\mu_\phi &\approx \phi(\mu) + \frac{1}{2}\sigma^2\phi''(\mu) \\ &= 2\arcsin\sqrt{\frac{k+0.5}{n+1}} + \frac{(k+0.5)^{-1/2}(n-k+0.5)^{-1/2}(2k-n)}{4(n+2)}\end{aligned} \tag{22}$$

where ϕ' and ϕ'' denote the first and second derivative with respect to ε. Finally, the asymptotic intrinsic q-credible interval in the reference parametrization is $[\phi_-, \phi_+]$ where

$$\phi_\pm \approx \mu_\phi \pm z_q\sigma_\phi = \mu_\phi \pm \frac{z_q}{\sqrt{n+2}} \tag{23}$$

where z_q is the $(q+1)/2$ quantile of the normal distribution. The intrinsic q-credible interval for the efficiency is obtained by transforming back to $\varepsilon_\pm = \varepsilon(\phi_\pm) = \sin^2(\phi_\pm/2)$.

The intrinsic q-credible intervals have also approximate q coverage when interpreted in the frequentist way. However please note that, depending on the actual values of the parameters n, k, the reference credible intervals can overcover or undercover the true value in repeated experiments, while the best practice in the frequentist approach is to choose intervals which never undercover (for more details, see [1]; a frequentist choice for the asymmetric confidence intervals is also reported on chapter 32 of [2]).

5. Non trivial use cases

So far, we assumed that all entries of the initial histogram had unit weight and had been selected by an independent binomial process. This may not be true in all cases, as it happens sometimes in high-energy physics. For example:

1. the initial histogram $\{(i, n_i)\}$ was obtained by scaling the simulated data sample to normalize it to some different value of the cross section. The histogram *should not be used* to make efficiency studies! Rather, the efficiency should be estimated by using the *original* histogram (filled with unit weights), in order to have a binomial process;
2. the initial histogram $\{(i, n_i)\}$ was obtained as the weighted average of several contributions (for example, by combining simulated samples corresponding to the same integrated luminosity but having very different cross sections). As above, the histogram *should not be used* to make efficiency studies, which require the use of the original histogram (filled with unit weights);
3. the initial histogram $\{(i, n_i)\}$ has been filled using weights ± 1, summing up terms which may give a positive or negative contribution to the final production probability. This is the case of the output from MC@NLO, and is addressed in section 5.1;
4. the numbers of events before and after the cut have been obtained with a different procedure than simply counting events. For example, they could come from fits as in section 5.2;
5. the initial sample was selected by a non independent process. This can be important when measuring the trigger efficiency and is addressed in section 5.3.

5.1. Events with positive or negative unit weights

In high-energy physics simulations, it might happen to work with samples filled with positive and negative unit weights, as it happens for example in the output of MC@NLO [12]. Each individual event is independently simulated, and knows nothing about its weight. Hence we can separately consider the samples with positive and negative unit weights, with n_+, n_- initial numbers of events and k_+, k_- entries after the selection. For each sample, the efficiencies ε_+ and ε_- can be computed individually following the methods already seen in previous sections: using the Jeffreys' prior, their posteriors are Beta distributions with parameters $a_i = k_i + 0.5$ and $b_i = n_i - k_i + 0.5$, with $i = +, -$.

However, we are interested in the overall efficiency, after subtraction of the two samples. For MC@NLO, its authors say that the efficiency should be estimated as $f = (k_+ - k_-)/(n_+ - n_-)$ when $k_+ \geq k_-$ or zero otherwise, and they suggest to use the usual "propagation of errors" to estimate its variance whenever the numbers are high enough that the binomial approximation holds, or to run many MC samples through the cuts and look at the dispersion in the result if the data sample is too small.

Here we use instead the method of moments to find the parameters of the posterior Beta distribution that matches the approximate mean $E(\varepsilon) \approx f$ and its approximate variance. One may write $f = (k_+ - k_-)/(n_+ - n_-) = (f_+ n_+ - f_- n_+)/(n_+ - n_-)$, which is our estimate for the weighted sum $(\varepsilon_+ n_+ - \varepsilon_- n_+)/(n_+ - n_-)$. The latter has variance

$$V(\varepsilon) = \frac{n_+^2 V(\varepsilon_+) - n_-^2 V(\varepsilon_-)}{(n_+ - n_-)^2} \tag{24}$$

where $V(\varepsilon_+), V(\varepsilon_-)$ are computed from the posterior PDFs of the individual samples with homogeneous weights. Putting $E(\varepsilon)$ and $V(\varepsilon)$ into equations (4) and (5) one can finally find the approximated posterior Beta density which gives the desired result. This is the recommended approach, given that the other method suggested below is much more complex.

Please note that the method above can be easily generalized to be used in case we are told to consider $f = (\sum_i w_i k_i)/(\sum_i w_i n_i)$ as the best estimate of the efficiency from a mixture of samples with weights w_i, provided that we know all pairs (n_i, k_i) of events before and after the selection for each homogeneous sample. In such case, we can apply the method of moments with $E(\varepsilon) = f$ and $V(\varepsilon) = [\sum_i w_i^2 V(\varepsilon)]/(\sum_i w_i)^2$ to find the approximated posterior PDF.

Another possibility is to use the general results obtained by by Pham-Gia et al. [13], who found a (rather complex) analytical expressions for the case in which one makes the difference between two independent random variables, each one following a Beta distribution. The relevant properties of their "beta-difference" distribution are summarized in appendix A.3 but have the following disadvantages with respect to the approximate solution explained above. First, in general the difference of two binomial parameters has domain in $[-1, +1]$, so that in our case the posterior needs to be set to zero for negative values and renormalized. A numerical approach needs to be used to handle the resulting posterior, equation (37) of appendix A.3. Second, Pham-Gia et al. [13] adopted the uniform prior, arguing that it can be considered non-informative. We have already explained in section 3.2 that this is not a proper choice, so that such solution is appropriate only when the uniform prior is justified as an informative prior PDF.

5.2. Events estimated by fits

Let us consider the case in which we have a distribution of N events which is a mixture of "signal" (S) and "background" (B) events, and we want to know the effect of some cut on the two components. We obtain separate estimates n_s, n_b for the initial numbers of S and B events by fitting the distribution with a function which is a mixture of two PDFs with relative weights x_s and $x_b = (1 - x_s)$, so that $n_s = x_s N$ and $n_b = (1 - x_s)N$ may be non-integer positive numbers with the constraint $n_s + n_b = N$. Later, we apply the cut under study and fit the resulting distribution, containing K events, to obtain the estimates k_s, k_b of the final numbers of S and B events. Again, the fits provides the relative weight y_s for the S component, so that $k_s = y_s K$ and $k_b = (1 - y_s)K$, with $k_s + k_b = K$. We are interested into the efficiency of the cut for signal events, which would be naively expected to be k_s/n_s, though this is not the correct guess, as shown below.

A possible way of approaching the problem is to consider the following distinct phases:

1. The first fit separates S from B events from the initial distribution and it is assumed to provide the best estimates of n_s and its RMS r_s. This can be considered a binomial process in which we select $n_s = x_s N$ signal candidates out of N, so that the result can be cast with the method of moments in the approximate form of a Beta density whose mean is x_s and whose variance σ_s^2 is also given by the fit: $\sigma_s = r_s/N$.

It is sufficient to use equations (4) and (5) with $E = x_s$ and $V = \sigma_s^2$. This means that we consider the weight x_s returned by the fit as the posterior estimate of the efficiency of the initial signal selection. The obtained beta density $\mathrm{Be}(x; a, b)$ will be used as prior PDF for the next step. Incidentally, we note that the non-integer nature of n_s and k_s is not a problem here, because the posterior Beta density is not required to have integer parameters.

2. We are interested into the cut efficiency for signal candidates, that is the probability that an event both passes the cut and is assigned to the S population by the second fit. We can either consider this a unique selection process or split it into $P(\text{cut, fit}) = P(\text{fit}|\text{cut})P(\text{cut})$. The latter case implies one more iteration of the Bayes' theorem, but here we consider the unique selection because it is simpler and we are not interested into the details of the second fit.

From the discussion above, we know that the posterior PDF describing the efficiency of our cut for signal candidates is a Beta density with parameters $a' = k_s + a$ and $b' = n_s - k_s + b$, where a, b have been determined with the method of moments from the first fit. Hence, the expected efficiency for our selection is given by equation (31) from appendix A.2: $a'/(a' + b') = (k_s + a)/(n_s + a + b)$, which is different from the naive value k_s/n_s, in that it explicitely takes into account the effects of the fitting procedure, i.e. of the classification in "signal" and "background" events.

Because in general $a \neq 1/2$ and $b \neq 1/2$, the final Beta density is different from what is obtained from simple event counting using Jeffreys' prior. Depending on the fit properties (expecially on the first fit), the final PDF may be wider than the reference posterior, most notably when the number of events is small. The fitting procedure has some cost: intuitively, a fit which better discriminates between signal and background events will provide a narrower density, whereas a poor fit will end up into a wider one. The good point of the Bayesian treatment is that it accounts for all the available information, not that it produces narrower distributions.

5.3. What to do if the samples are not independent?

The case in which the initial histogram $\{(i, n_i)\}$ does not represent a statistically independent sample is expecially important in trigger efficiency measurements, when there is no other trigger selection which is statistically uncorrelated with respect to the signature A under study. Figure 3 shows

$$\begin{tikzcd} \boxed{0} \arrow[r, "B"] \arrow[d, "A"'] & \boxed{2} \arrow[d, "A"'] \\ \boxed{1} & \boxed{3} \arrow[l, "???"'] \end{tikzcd}$$

Figure 3: Non-independent selection. The sample $\boxed{0}$ only exists with simulated data, that can be used to make a first estimate of the efficiency of selection A by taking the bin-wise histogram division $\boxed{1}/\boxed{0}$. With real data, the least biased sample $\boxed{2}$ is obtained by selecting events with B. Next, imposing condition A on $\boxed{2}$ one gets the sample $\boxed{3}$.

a situation in which one wants to study the systematic effect of a previous trigger selection B on A starting with the true distribution $\boxed{0}$, on MC data. As we have seen, the best estimate of the efficiency of A as function of some quantity x is given by the histogram ratio between the distribution of x after the selection (the histogram filled with sample $\boxed{1}$) and its distribution before (sample $\boxed{0}$). Real data can only be taken with trigger B (which in practice is chosen to be the least correlated as possible to A), obtaining sample $\boxed{2}$. Later, condition A can be required on $\boxed{2}$ obtaining sample $\boxed{3}$, which has been selected by requiring *both* B and A, and the histogram division $\boxed{3}/\boxed{2}$ estimates the probability $P(A|B)$ to select one event with A, given that it was already selected by B.

In order to find the desired true efficiency $P(A)$, we make use of the relation defining the conditional probability, $P(A \cdot B) = P(A|B)P(B) = P(B|A)P(A)$, obtaining $P(A) = P(A|B)[P(B)/P(B|A)]$, where the fraction in brackets cannot be determined with real data alone. The true efficiency of B alone is found with simulated data by requiring condition B on sample $\boxed{0}$, whereas the relative efficiency $P(B|A)$ of B with respect to A is obtained by requiring condition B on sample $\boxed{1}$. The conclusion is that, without some statistically independent trigger, one can not estimate the trigger efficiency using real data only. Rather, a simulation is required to measure the impact of the non-independent trigger B on the selection A under study.

If the approximation in which A and B are independent is good enough (MC data can be used to check this), the value of the fraction can be considered equal to one in all bins, and the bin-wise ratio between $\boxed{3}$ and $\boxed{2}$ gives a good estimate of the true A efficiency. Such approximation is justified if the (systematic) effect of $P(B)/P(B|A)$ is small compared to the statistical uncertainty on the ratio between $\boxed{3}$ and $\boxed{2}$ (which might not be true in all bins), i.e. when the square root of the variance is significantly larger than $1 - P(B)/P(B|A)$. Otherwise, the recommended approach is to model the knowledge about such ratio with a Beta density in each bin, and obtain the

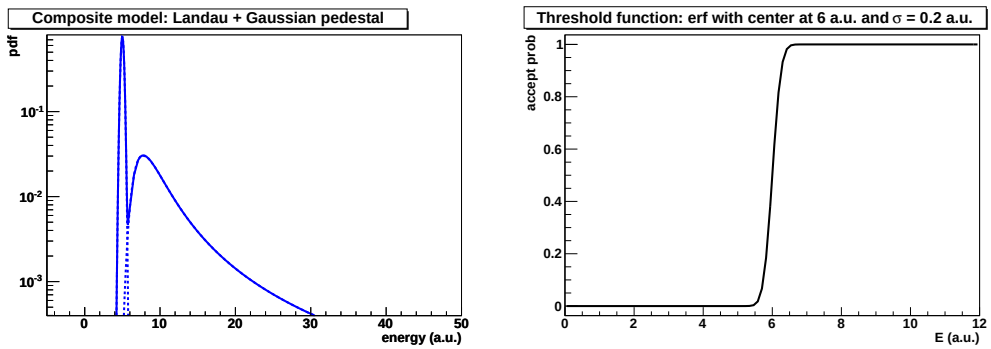

Figure 4: Probability density function followed by the simulated sample (left) and true threshold efficiency (right) as function of the energy (in arbitrary units). The Landau distribution has parameter = 8.0 and scale = 1.0 whereas the Gaussian pedestal peaks at 5.0 a.u. with 0.2 a.u. standard deviation. The threshold (right) is modeled as an error function with center at 6.0 a.u. and standard deviation of 0.2 a.u.

Beta posterior for the efficiency of the selection A as explained before, using the result from the MC encoded in the form of the prior Beta density.

6. Fitting examples

Here we consider a "real life" example: a simulation of a common experimental setup, the measurement of the selection efficiency versus the threshold on a scalar quantity, and the fit of the resulting "turn-on" plot in ROOT. Because no Bayesian fitting technique is yet available in ROOT[7], a pragmatic approach is to test the different fitting options available in the release (5.26/00) used for this work.

The model describes the energy lost by minimum ionizing particles (MIPs) while crossing a thin slab of active material (e.g. a scintillator) as a Landau distribution. It is assumed that the read-out electronics (e.g. a photomultiplier tube read by a charge integrator) is tuned to have a dynamic range large enough that the peak of the MIP energy distribution is not very distant from the pedestal (figure 4, left panel; the energy is in arbitrary units, e.g. ADC counts). The experimental setup is triggered by a comparator whose threshold is somewhere in between the two peaks. Due to the electronic jitter, the comparator does not apply a sharp cut on the distribution. Rather,

[7]Work is in progress on this side. Lorenzo Moneta, private communication.

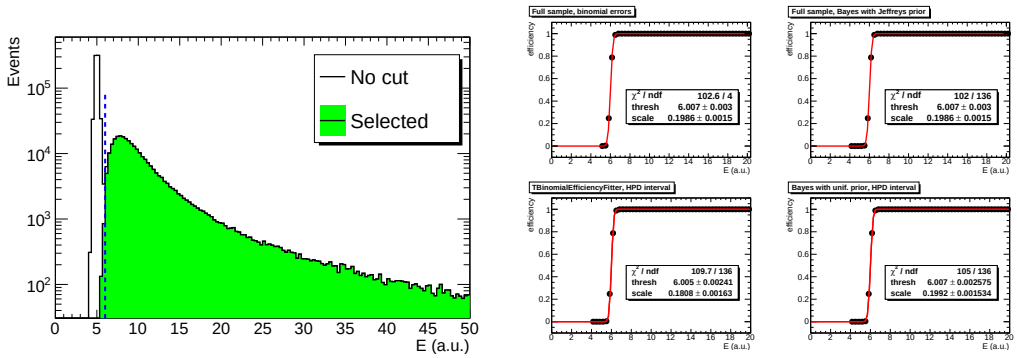

Figure 5: Left: full spectrum, before and after the cut (the threshold position is indicated by the dashed line). Right: fits done with binomial errors (top-left), the square root of the reference variance (8) (top-right), HPD credible intervals (bottom-right), and TBinomialEfficiencyFitter (bottom-left).

a smooth "turn-on" function (figure 4, right panel) is obtained, due to the random fluctuations on the difference between the measured energy and the threshold. The ideal (and quite common) case of Gaussian fluctuations has been simulated, so that the turn-on curve is an "error function" (the Gaussian integral from minus infinity to the considered value).

A total of 1 million events has been simulated, with energy following the distribution shown in the left plot of figure 4. Each event has been "rejected" accordingly to the true threshold function shown in the right plot of the same figure[8]. Later, the best estimate of the efficiency has been obtained by taking the bin-wise ratio of the energy histograms filled for the events passing the threshold and for all events.

Figure 5 shows the full sample, before and after the cut, and four different fits performed with the following function:

$$g(x) = b + \frac{p}{2}\left[1 + \operatorname{erf}\left(\frac{x-t}{\sqrt{2}\,w}\right)\right] \tag{25}$$

where t is the best estimate of the threshold position, w the threshold width (Gaussian standard deviation), p is the plateau value reached at high energy (left fixed at 1 in the fits) and b is the lowest efficiency (i.e. the "offset" of

[8]The event was rejected when the value returned by a uniform generator was higher than the threshold function computed at the event energy. Both energy and the decision have been saved on file for later use.

the "turn-on" curve, left fixed at zero).

One of the fits is performed using the frequentist approach, implemented in the ROOT class TBinomialEfficiencyFitter, which fits the experimental points with a theoretical model using the maximum likelihood method (the input histograms must be filled with weights of 1). The best values are found by maximizing the sum of the binomial log-likelihoods defined for each bin in the input histograms (which must have the same binning) and, from the Bayesian point of view, it gives the correct answer when the binomial approximation of section 2.2 is acceptable, i.e. when both k, n are sufficiently large, a quite usual case.

As expected with so many events, the fit results using the binomial approximation (ROOT option `"B"` of TH1::Divide()), the standard deviations computed with the Jeffreys' prior, the asymmetric credible intervals recommended by Paterno, or TBinomialEfficiencyFitter agree with each other. The latter method underestimates the threshold jitter obtaining a result which is not compatible with the true value (0.2 a.u.) within the quoted uncertainty, whereas the values from the other fits are at about 2 standard deviations from the true model.

Apart from TBinomialEfficiencyFitter, which makes use of the two histograms filled before and after the cut, the other fits use the value of the ratio in each bin and the uncertainty which we have assigned following different methods. Such fits have been made with the option `"ME"` of TH1::Fit() to select error estimation using the Minos technique and the improved fit results by TMinuit. Another possible approach is to use the option `"LL"` to use the log-likelihood method instead of the chi-square method, when the bin contents are not integer values — such option is not supported by TGraph::Fit() hence was not used when fitting the output of TGraphAsymmErrors::BayesDivide() —. In this case, the estimated errors are much larger than with `"ME"` (also, the reported quality of the fit is worse) but again the results are compatible with the true model.

Next, the sample has been divided into 100 subsets of 100, 1000, and 10000 events each, and repeated turn-on fits have been attempted with all methods. The values of the threshold and width from all fits have been histogrammed (only when the fit was successful) as shown in the figures 6 and 7, in which the fit options `"ME"` and `"LL"` are compared for the "binomial errors" (option `"B"` for the histogram division) and the reference standard deviations, and in figure 8, showing the results by TBinomialEfficiencyFitter and by the chi-square fit of the graph showing Paterno's HPD credible

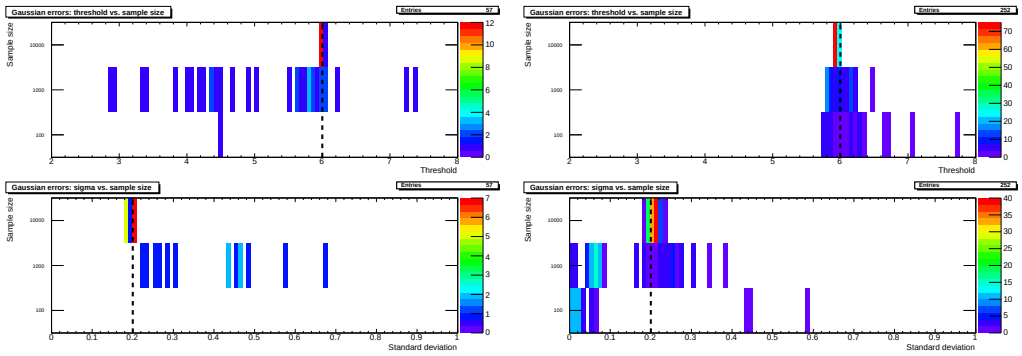

Figure 6: Distribution of the fit parameters obtained with the binomial errors. Chi-square (left) and log-likelihood (right) fitting methods are compared.

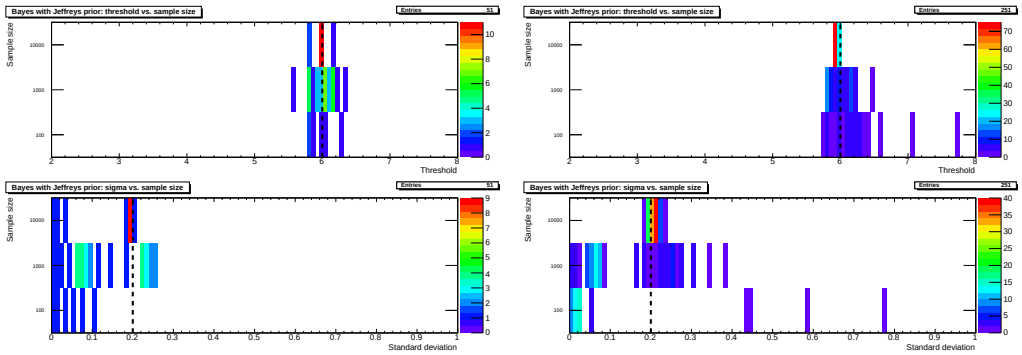

Figure 7: Distribution of the fit parameters obtained with the square root of the reference variance (8). Chi-square (left) and log-likelihood (right) fitting methods are compared.

intervals.

These figures show that the log-likelihood fit (option `"LL"`) obtains results which are more closely clustered around the true values when using the binomial approximation for the errors (figure 6). However, as noticed above, this approximation is very bad when one has most points at zero or full efficiency, because it assigns zero uncertainties to such cases. This problem is visible also with the full sample: in figure 5 the fit with binomial errors has a very low number of degrees of freedom, compared to the other three cases (bins with zero uncertainty are not counted as degrees of freedom). When using the option `"LL"` the number of degrees of freedom increases to 148, so that this is preferrable when using the approximation of binomial errors.

In our example, the turn-on is so sharp that, expecially for small samples, is is very likely that there is no single bin at intermediate efficiency, which

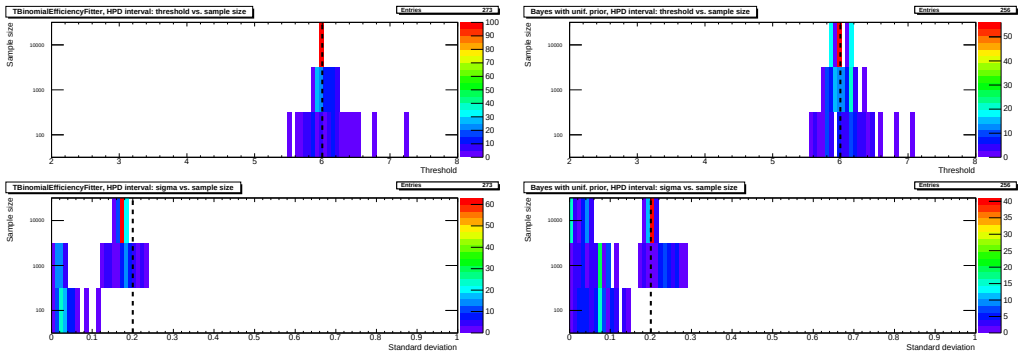

Figure 8: Distribution of the fit parameters obtained with TBinomialEfficiencyFitter (left) and HPD credible intervals (right).

makes the fit fail in most cases with the option `"ME"`, though in less cases with the option `"LL"`. Even with larger samples, the passage from zero to full efficiency happens in a couple of bins only, so that this is admittedly a quite pathological situation. Having no bin at intermediate efficiencies is a problem for all fitting techniques, though TBinomialEfficiencyFitter seems to suffer less about it. Its clustering around the true values is acceptable and similar to the fit done with Paterno's HPD credible intervals (figure 8), but the latter fails many more times. When using the standard deviations computed with the Jeffreys' prior, maximizing the log-likelihood instead of minimizing the chi-square helps reducing the number of failing fits too, though the clustering around the true values does not improve significantly.

In conclusion, TBinomialEfficiencyFitter is the most robust way of fitting the efficiency implemented in ROOT, though in our test it appears to underestimate the width of the turn-on curve. On the other hand, the binomial approximation is by far the worst approach. If the samples are large enough that it is safe to use the binomial approximation for fitting, then one should always use the option `"LL"`, which in our test improves the clustering around the true values. Even with such option, the performance is similar to what is obtained by the fits based on the reference standard deviations (but the latter are to be certainly preferred if the default chi-square fitting method is used). Finally, though it is possible to directly fit the graph showing Paterno's HPD credible intervals, there is no advantage in doing so. The best approach is to fit the data using another method and plot the resulting function on top of the graph showing the HPD credible intervals, which are the only asymmetric efficiency intervals already available in ROOT, though

one could implement other things[9].

7. Summary

Estimating the selection efficiency is a fundamental task in most data analyses, based on simulated and/or real data. The measured relative frequency provides the best estimate of the true efficiency in the frequentist approach and coincides with the posterior mode obtained in the Bayesian treatment with uniform prior. However, such prior cannot be considered non-informative. Instead, if we are completely uncertain about the efficiency before making the experiment, the use of the Jeffreys' prior is recommended.

In general, if some prior knowledge is available, it is recommended to encode it into a function belonging to the family of Beta distributions, whose parameters can be determined with the method of moments if an approximation is needed. This ensures that the posterior also belongs to the same family, so that all properties summarized in appendix A.2 are immediately available. An important example of the use of informative priors is the combination of independent samples, which is also the correct way for including prior knowledge coming from simulations to model systematic effects.

The knowledge of the uncertainty on the efficiency is needed when scaling observed quantities to estimate their original values (e.g. the true rate). In this case, the recommended approach is to use the mean and variance of the posterior PDF in the computation, whenever the use of the full posterior is not practical. The usual variance algebra holds, with the caveat that the square root of the final variance might not be good to define a symmetric credible interval, because of the inherent asymmetry of the posterior in the general case. Though in many applications the posterior will be significantly peaked around the true value, so that the binomial (symmetric) approximation holds, care needs to be taken when handling very low or very high efficiencies, and when the number of events is relatively small, because such approximation behaves poorly in such cases (using the full posterior is always better, if possible).

When communicating the result of an efficiency measurement, the recommended approach is to provide the $2 + 2$ Beta parameters corresponding to the posterior and prior PDFs, so that the user will be able to test the effects

[9]For coherence, it would be best to couple TBinomialEfficiencyFitter with the 95% confidence intervals defined on chapter 32 of [2].

of different priors and to combine the posterior with other independent measurements. In the plots, the observed frequency should be accompanied by asymmetric error bars. Paterno's HPD credible intervals are already available in ROOT via TGraphAsymmErrors::BayesDivide() but are not necessarily the best choice: reference credible intervals would be a better option, as shown in section 4.1.

When fitting the efficiency with a theoretical function, few different methods are available in ROOT. As it is shown in section 6, it is important to avoid performing a chi-square fit using binomial errors (which unfortunately seems to be the easiest choice, though it is the worst solution). The TBinomialEfficiencyFitter method is the most robust way of fitting turn-on plots in ROOT, though a better parameter estimation is obtained by performing a fit with the reference standard deviations (it is suggested to try using the option `"ME"` first, and switch to `"LL"` in case of failure).

Finally, special care must be used when handling samples that do not have unit weights or are not independent. Few recipes to deal with the most common use cases in particle physics have been sketched in section 5.

A. Useful relations

This appendix summarizes mathematical definitions and properties that are useful to deal with binomial processes. They can be found in standard books like [14].

A.1. Gamma function

The *Gamma function* is defined on the complex plane ($z \in \mathbb{C}$):

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, \mathrm{d}t \tag{26}$$

with $\Gamma(z+1) = z\,\Gamma(z)$. For integer values, $\Gamma(n) = (n-1)!$.

A.2. Beta distribution

The Euler *Beta function* is a symmetric function of $a, b \in \mathbb{R}$:

$$B(a,b) \equiv \int_0^1 t^{a-1}(1-t)^{b-1} \, \mathrm{d}t = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)} = B(b,a) \tag{27}$$

and the *incomplete Beta function* is

$$B_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t\ . \tag{28}$$

with $x \in [0,1]$.

For $x \in [0,1]$, the *Beta distribution* has PDF

$$f(x;a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \equiv \mathrm{Be}(x;a,b) \tag{29}$$

and cumulative distribution function

$$F(x;a,b) = \int_0^x f(t;a,b)\,\mathrm{d}t = \frac{B_x(a,b)}{B(a,b)} \equiv I_x(a,b) \tag{30}$$

where $I_x(a,b) = 1 - I_{1-x}(a,b)$ is the *regularized incomplete Beta function.* The mean E, mode m, variance V and skewness γ_1 of the Beta distribution (29) are

$$E(x;a,b) = \frac{a}{a+b} \tag{31}$$

$$m(x;a,b) = \frac{a-1}{a+b-2} \tag{32}$$

$$V(x;a,b) = \frac{ab}{(a+b)^2\,(a+b+1)} \tag{33}$$

$$\gamma_1(x;a,b) = \frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}} \tag{34}$$

Finally, the characteristic function is

$$\phi(t) = \int_0^1 \mathrm{Be}(x;a,b)\exp(-2\pi i x t)\,\mathrm{d}x = {}_1F_1(a;a+b;it) \tag{35}$$

where ${}_1F_1(a;b;c)$ is the confluent hypergeometric function of the first kind.

A.3. Posterior density for the difference

Here we consider two random variables that correspond to the selection efficiencies ε_+ and ε_- for the samples with positive and negative weights considered in section 5.1. We want to find the posterior for the difference $\varepsilon = \varepsilon_+ - \varepsilon_-$ using the general result found by Pham-Gia et al. [13]. Their

expression is valid for the general difference of two Beta-distributed random variables, with domain ranging from -1 to $+1$. However, we know that the physical efficiency can not be negative, hence we restrict the posterior to $[0, 1]$ (the normalization needs to be recomputed).

They chose to use a uniform prior so that their posteriors, having counted n_+ and n_- initial events and k_+ and k_- entries after the selection, are Beta distributions with parameters $a_+ = k_+ + 1$, $b_+ = n_+ - k_+ + 1$ and $a_- = k_- + 1$, $b_- = n_- - k_- + 1$. The correspondence between their and our notation, when dealing with the posterior under the assumption of uniform priors for ε_+ and ε_-, is: $\alpha_1 \equiv 1 + k_+$, $\alpha_2 \equiv 1 + k_-$, $\beta_1 \equiv 1 + n_+ - k_+$, $\beta_2 \equiv 1 + n_- - k_-$.

Their result (equations (2a) and (2c) in [13]) can be rewritten in our case (being $\alpha_1 \geq 1$) in a more compact form, which makes use of the third Appell hypergeometric function

$$\begin{aligned} F_3(a, b, c, d; e; x, y) = \\ = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{(a)_m \, (b)_n \, (c)_m \, (d)_n}{(e)_{m+n}} \, \frac{x^m}{m!} \, \frac{y^n}{n!} \end{aligned} \tag{36}$$

obtaining an expression valid for $0 \leq \varepsilon \leq 1$ (to be renormalized):

$$\begin{aligned} f(\varepsilon) \propto \frac{f(\varepsilon; \alpha_2, \beta_1)}{f(\varepsilon; \alpha_1, \beta_1) \, f(\varepsilon; \alpha_2, \beta_2)} (1 - \varepsilon)^{\alpha_2 + \beta_1 - 1} \times \\ \times F_3(\beta_1, \, \alpha_2, \, 1 - \alpha_1, \, 1 - \beta_2; \, \alpha_2 + \beta_1; \, 1 - \varepsilon, \, 1 - \varepsilon) \, . \end{aligned} \tag{37}$$

References

[1] R.D. Cousins, K.E. Hymes, J. Tucker, "Frequentist Evaluation of Intervals Estimated for a Binomial Parameter and for the Ratio of Poisson Means", NIM A 612 (2010) 388–398, arXiv:0905.3831.

[2] C. Amsler et al., "Review of Particle Physics", Phys. Lett. B667 (2008) 1.

[3] F. James, "Statistical Methods in Experimental Physics: 2nd Edition", World Scientific, 2006.

[4] I. Antcheva et al., "ROOT — A C++ Framework for Petabyte Data Storage, Statistical Analysis and Visualization", Subm. to Comp. Phys. Comm. 40th Anniversary Issue, 2009. `http://root.cern.ch`.

[5] M. Paterno, "Calculating Efficiencies and Their Uncertainties", FERMILAB-TM-2286-CD, May 22, 2004.

[6] T. Ullrich and Z. Xu, "Treatment of Errors in Efficiency Calculations", physics/0701199, 2007.

[7] J.M. Bernardo, "Reference analysis", Handbook of Statistics 25 (D.K. Dey and C.R. Rao eds.). Amsterdam: Elsevier (2005) 17–90.

[8] J.O. Berger, J.M. Bernardo, D. Sun, "The formal definition of reference priors", Annals of Statistics 37 (2009) 905–938.

[9] J.M. Bernardo, "Modern Bayesian Inference: Foundations and Objective Methods", Philosophy of Statistics (P. Bandyopadhyay and M. Forster, eds.), Amsterdam: Elsevier, 2009.

[10] H. Jeffreys, "An invariant form for the prior probability in estimation problems", Proc. Royal Soc. London A Math. and Phys. Sci., vol. 186, no. 1007 (1946) 453–461.

[11] J.M. Bernardo, "Intrinsic credible regions: An objective Bayesian approach to interval estimation", Test 14 (2005) 317–384.

[12] S. Frixione and B.R. Webber, "Matching NLO QCD computations and parton shower simulations", JHEP 06 (2002) 029; arXiv:hep-ph/0204244.

[13] T. Pham-Gia, N. Turkkan, P. Eng, "Bayesian analysis of the difference of two proportions", Comm. Statist.—Theory Meth., 22:6 (1993) 1755–1771.

[14] Abramowitz, M. and Stegun, I. A. (Eds.), "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables", 10-th printing. New York: Dover, 1972.