arXiv:0908.0130v5 [physics.data-an] 8 Sep 2010

physics 0908.0130

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 framework) 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.

---

[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.

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 $\varepsilon \to f$:

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

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

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

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