Analysis of multichannel measurements of rare processes with uncertain expected background and acceptance

I. B. Smirnov*

Petersburg Nuclear Physics Institute, Gatchina 188300, Russia

Abstract

A typical experiment in high energy physics is considered. The result of the experiment is assumed to be a histogram consisting of bins or channels with numbers of corresponding registered events. The expected background and expected signal shape or acceptance for a certain full signal rate are measured in a separated auxiliary experiments, or calculated by the Monte Carlo method with finite sample size, and hence with finite precision. An especially complex situation occurs when the expected background in some of the channels happens to be zero in the auxiliary background experiment due to either a fluctuation of the auxiliary measurement or because it is truly zero. Different statistical methods give different confidence intervals for the full signal rate and the different significances of the signal+background hypothesis versus the pure background hypothesis.

1 Introduction

Rates of rare processes in high energy physics have sometimes to be estimated from a few observed events. This can happen at the research of very rare processes or at the beginning of any research. Reconstruction of such rates is a complicated problem with fundamental ambiguity [1], especially in the presence of the uncertain nuisance parameters.

1.1 Typical experiment

The result of an experiment is frequently represented by a histogram consisting of several, \( k \), bins or channels, \( k \geq 1 \). Each of these channels keeps the number of events \( n_i \) registered in this channel, where \( i \) is the number of channel. This number is sampled from the Poisson distribution with a certain parameter, which is unique for each channel. Events in each channel are expected to be produced by background processes, called background, and by a studied process called a signal, all distributed according to the Poisson law. The expected background \( b_i \) and expected signal or acceptance \( a_i \) (we will prefer the term “expected signal”) are either known precisely or measured in separated auxiliary experiments, or calculated by the Monte Carlo method with finite statistics, and hence with a finite precision. In general case they can correspond to different expositions or luminosities, so the expected full rate in the main experiment is expressed by

\[
f_i = t_a a_i s + t_b b_i ,
\]

or in the vector notation,

\[
\vec{f} = t_a \vec{a} s + t_b \vec{b} ,
\]

where \( t_a \) and \( t_b \) are the ratios of expositions of the main and respective auxiliary experiments and \( s \) is the real signal rate, the absolute value or the value relative to the expected signal rate.

We will consider here only the stochastic uncertainties of \( a_i \) and \( b_i \). The uncertainties that are presumably non-stochastic can usually be assumed stochastic in some more generic sense and can be handled by similar methods. Both \( a_i \) and \( b_i \) are assumed to be “measured” in the respective “auxiliary experiments” as the numbers \( n_{ai} \) and \( n_{bi} \) sampled from the Poisson distributions with the corresponding parameters \( a_i \) or \( b_i \).

The task of the experimental research is to determine the most probable full signal rate, a confidence interval for it, and the significance of the signal (plus background) hypothesis versus the pure background hypothesis.

1.2 Zero channels of expected background

An especially challenging situation arises when the expected background in some of the channels happens to be zero in the auxiliary background experiment due to either a fluctuation of the auxiliary measurement or because it is truly zero. Such situation can happen at the searches of very rare processes in experiments with very good background rejection, for which it is difficult to perform Monte Carlo simulation of background with large final statistics, because in order to do this it needs to run the huge initial statistics. Hence it is not possible to distinguish the
cases of a downward background fluctuation and the true zero background on the base of existing information. It is then unclear both conceptually and numerically, how to interpret non-zero result of the main experiment in this channel. The opposite case of very small acceptance and zero expected signal channels does not seem similarly confusing, if there is non-zero expected background for this channel.

Literature does not offer any certain receipt to deal with zeros in the expected background.

One could simply ignore such channels. Otherwise, zeros can be removed by unification or smoothing of neighboring or presumably similar channels, some of which are non-zero. The similarity of the channels is typically indicated by neighboring values of a response variable produced by multivariate analysis methods. But any zero-removing procedure can lead, briefly speaking, to unexpected change of the precision and complicates its estimation. The most doubtful case is when a few zero channels appear in the distribution of the expected background by the response variable at the end of the distribution, where the expected signal is maximal. This situation is the most probable, since mainstream multivariate analysis methods, such as artificial neural networks, always produce the reducing expected-background histogram and increasing expected-signal histogram. In this case, the end of the spectra, which has to be the most important region for the results, is subjected to effectively arbitrary treatment.

Therefore, it is interesting to investigate which statistical methods, if any, provide correct results in such tasks, when the channels content is taken as is, without forbidding of zeros or smoothing.

1.3 Tests of intervals

The idea is to generate a sequence of pseudo-experiments with some small enough true background, which ensures a large probability of zero content in the bins of the expected-background histogram, when the latter are divided into many channels. Then to divide the experimental results into different number of channels and to treat them by all available statistical methods. Then one can check and compare their various actual characteristics such as the size of confidence intervals for different divisions, the probability of covering the true known parameter by the interval reconstructed by given method for given division, possibility and probability of producing meaningless intervals. Methods usually exhibit different characteristics for different divisions. It is interesting to see whether an optimal number of channels can be chosen on the base of the data available for the experimenter, for instance, by minimizing the interval width reconstructed by the data of the single experiment (not the average width). The danger is that too selective optimization could lead to fictitious improvement. Examples of fictitious improvement are the “flip-flopping” effect or the known “look-elsewhere” effect. A rude optimization with choice between “5 or 10” channels with equal width by the response variable or “2 or 5” ones etc., using the change by the factor around 2 or 3, looks reasonable and is the subject of this research. For each next identical experiment with different fluctuated observations one can obtain different optimal division. Since the true value of the parameter of interest is known, we can check whether the correct coverage is provided both for fixed and for optimized divisions.

This is effectively a “frequentist” or a “classical” viewpoint. The both terms are accepted, see, for example, Refs. [1], §26.1 of [3, 4]. If the obtained interval includes (“covers”) the true value, whatever it is, with the stated probability, or frequency in a long sequence of experiments, it is said that the coverage is provided. Then, from the formal classical viewpoint the method is acceptable. During the generation of experiments the nuisance parameters are assumed to be fixed at their true values, whatever they are, too. At the analysis they are unknown, but they are not requested to be reconstructed or “covered”.

Such an investigation is interesting not only in the context of the problem with zeros, but in a much wider context. The most useful methods provide the guaranteed coverage at the precisely known expected background and signal shape, but do not guarantee it if the expected background and signal are known only approximately. The Bayesian method and the profile likelihood (or likelihood ratio) method do not guarantee the frequentist coverage at all [1], but it is interesting to see whether they provide it in practice. If the parameter of interest is restricted, typically to be non-negative, some classical methods can produce empty or unphysically small intervals at downward fluctuation of background and small true signal (c.f. Refs. [2, 5]). At the full Neyman construction by all parameters including the nuisance ones the projection of the confidence set to the parameter of interest will usually “badly over-cover” [6]. One else issue which is discussed in the literature is the coverage of the upper border of the classical intervals for the experiments microscopically dependent on the signal [7].

Frequentist tests reported in Refs. [8, 9, 10, 11] do not include some interesting methods and cases, such as, in particular, zeros in the expected background.

1.4 Tests of significance

We can test the interval-finding method by checking the coverage of intervals it produces. However, it is not so simple to formulate a similar test for significance. The probability that background imitates the observed or larger signal should naturally depend on the true background. If the background is only assumed, this probability, as a rule, can only be assumed too. It is easy to calculate the latter for given assumption, but it will usually differ for different assumptions. Which of its values is true?

Some works, see for example [12, 13], test whether the p-value distribution is uniform in the precise or conservative integrated sense. Denoting the p-value by \( \rho \) and its probability density distribution by \( p(\rho) \) one can check whether \( \int_0^{\rho_1} p(\rho) \, d\rho \leq \rho_1 \) for any small enough threshold
If the sign is “<”, this is the conservative case, better than assumed to be necessary. For example, in the work [12] the values $\rho_t$ corresponding to significances 1.28, 3, and 5 are tested. Actually the work [12] tests significances but the difference is unimportant due to the one-to-one correspondence.

However, the feature of uniformity, necessarily satisfied for the known nuisance parameters, is not sufficient for the case of unknown nuisance parameters. Even if the conservative case occurs, the corresponding $p$-values can have no relation to reality. Furthermore, although it is hard to imagine good or reasonable $p$-values with the integral greater than $\rho_t$, since this seems to indicate the systematically too optimistic estimates of significance, there seems to be no rigorous proof of that.

The integrated uniformity is necessary and sufficient for the case of known nuisance parameters since greater test statistic always corresponds to lower $p$-value. This is so because the latter is obtained by the integration of the test statistic distribution. However, in the case of unknown nuisance parameters the true test statistic distribution is unknown and the $p$-value is obtained differently. Both the test statistic and the estimate of $p$-value by any concrete method depend in a non-trivial way on the auxiliary measurements. The test statistic and the $p$-value are not obliged to vary synchronously. Some of the methods described later were specially checked and the pathological couples of events with the simultaneous rise of both the test statistic and the $p$-value were found for many-channel cases. Hence, the integrated uniformity is not sufficient, at least. The $p$-value distribution can be uniform, but $p$-values can nevertheless be wrong from some point of view or according to some test. If the $p$-value distribution is uniform, one can, for instance, replace $p$ by $1 - p$ or do more selective exchanges and obtain the uniform distribution as well. If the uniformity is deemed to be sufficient, both distributions will be formally correct, but at least one of them (if not both) will not have any relation to the real $p$-values.

Finally, the uniformity test can be impossible because of large computer time requirements.

Noticing that it is easy to test intervals but difficult to test the significance one can ask whether the method that provides intervals with coverage provides also the good significance. This paper will show that this is usually so.

The particular case of the one-channel “on/off” problem solved through the test of the ratio of Poisson means [12] yields the significance that appears to be independent on the background hypothesis, though dependent on the total $n_{on} + n_{off}$ measurement. However, there is no similar simplification known for the many-channel case.

Ref. [14] has suggested that there exists “a fully frequentist method for hypothesis testing” with a Neyman construction in each of the nuisance parameters, their corresponding auxiliary measurements and a likelihood ratio test statistic. However, Ref. [14] has briefly mentioned some “subtleties”. Moreover, it is very difficult, if at all possible [6], to construct the confidence belt practically, for example, for 60 parameters and 61 measurements, as would be necessary for a simple test example considered in this paper.

1.5 Content

This paper is organized as follows. In the next section the test problem is described in more detail. The optimization of divisions is briefly outlined.

The following main groups of methods are described and tested in the following sections:

- The Bayesian approach (section 3) [1, 14, 15]. In order to improve the frequentist coverage of the credible intervals this section introduces a special “safe” priors for treating the nuisance parameter uncertainties and proposes a special modification of the central intervals.

- The frequentist (classical) treatment of the maximum likelihood estimate (section 4). Descriptions without the nuisance parameters are in Refs. [5, 19]. After introduction of the method this section describes various variants of their inclusion and the results of the tests. The last subsection of this section discusses the connection between the coverage of intervals and correctness of significance. It also introduces new frequentist-style tests of significance.

- The profile likelihood or the likelihood ratio method (section 5) — currently both notations are used in the literature, see Refs. [1, 10, 14, 15].

- The $CL_s$ methods (section 6). The frequentist treatment of likelihood ratios with correction for microscopic signal dependency. Descriptions of the background-related method without the nuisance parameters are in Refs. [7, 20, 21]. The new maximum-related method, see Refs. [22, 23], includes the nuisance parameters uncertainties. Its asymptotic approximation is useful for quick estimations.

Section 7 provides a comparison of significances obtained for the one-channel “on/off” problem by different methods for particular cases compiled in Ref. [12].

The last section briefly summarizes the main features of all methods.

2 The test problem

2.1 Notations

The conditional probability (density) of observing an experimental output $x$ at given parameter $y$ is denoted by $P(x|y)$ for the discrete case and by $p(x|y)$ for the continuous case, except for the Bayesian prior distributions.

The calculations on which this paper is based were performed by a self-made C++ software package independent on the external statistical software. To my knowledge, the conclusions of this research paper do not currently represent official recommendations of any collaboration or institute.
which are denoted by $\pi(x)$. Both $P$ and $p$ denote probabilities (densities for $p$) of observing corresponding values, but not functions of a fixed form. We remain the same notations even when we consider them as functions of $y$ and call “likelihoods” ($38.22$ in Ref. [24] and Ref. [1]).

Then, for instance, the joint probability of obtaining $n_i$ events in $i$-th channel of the main experiment, $n_{ai}$ in the auxiliary signal experiment, and $n_{bi}$ in the auxiliary background experiment is denoted by

$$P(n_i, n_{ai}, n_{bi}|s, a_i, b_i) = P(n_i|s, a_i, b_i)P(n_{ai}|a_i)P(n_{bi}|b_i) =$$

$$P(n_i|t_aoe + t_boe)P(n_{ai}|a_i)P(n_{bi}|b_i). \quad (3)$$

If all channels are involved, the corresponding multiplication of probabilities is denoted by

$$\prod_i P(n_i, n_{ai}, n_{bi}|s, a_i, b_i) = P(\vec{n}, \vec{n}_{a}, \vec{n}_{b}|s, \vec{a}, \vec{b}) =$$

$$P(\vec{n}|t_aoe + t_boe)P(\vec{n}_{a}|\vec{a})P(\vec{n}_{b}|\vec{b}). \quad (4)$$

The elementary probabilities $P(n|\mu)$ of observing $n$ events with the average expectation $\mu$ for this work are assumed to follow the Poisson law:

$$P(n|\mu) = \text{Poisson}(n, \mu) = \frac{\mu^n e^{-\mu}}{n!}. \quad (5)$$

### 2.2 Parameters and algorithm

In this paper it will be assumed that the multivariate response variable, denoted $x$, is varied from 0 to 1. This interval is equally divided into $k$ bins with step $1/k$ and end points $x_i$, $x_{i+1}$. The true background and signal distributions are

$$f_b(x) = Ce^{-a_x}, \quad f_a(x) = Ce^{-q(1-x)}$$

respectively, where the normalization factor $C = \frac{(1 - e^{-q})}{q}$ is needed to make the integral equal to unity. The value of $q$ is taken equal to 3 in this paper. The true parameters $a_i$ are determined by equalities:

$$a_i = N_a \int_{x_i}^{x_{i+1}} f_a(x) \, dx, \quad b_i = N_b \int_{x_i}^{x_{i+1}} f_b(x) \, dx, \quad (6)$$

where $N_a$ and $N_b$ are the mean total numbers of detected events in the corresponding auxiliary experiments.

In order to generate the auxiliary pseudo-experiments one can either generate the numbers of actual signal events by the Poisson law with the mean $N_a$ and similarly for the background events with $N_b$ and distribute these events randomly according to Eq. (3), or to generate the numbers in each bin by the Poisson law according to the means given by Eq. (6). Two separate histograms, one with expected signal and the other with expected background, are filled for each couple of auxiliary pseudo-experiments. If the case of precisely known parameters is considered, the corresponding histograms can be filled by $a_i$ or $b_i$ or both, depending on the case. Similarly in the simulated main pseudo-experiment the events are generated by either of two ways with taking into account Eqs. (6) and Eq. (11) with $s = s_{\text{true}}$, the true signal rate, which is unknown for the analysis program and has to be reconstructed by it. To obtain identical events with different number of channels only the histograms with the largest number of channels are filled by the method described above. The other ones are obtained by summing up the content of the neighboring channels. We will consider divisions with 1, 2, 3, 5, 10 and 30 channels. This is all done in a separate “main” program. For each division the main program calls the analysis program and transmits to the latter three histograms: the “main”, and two “auxiliaries”.

The analysis program knows $t_a$ and $t_b$ and reconstructs the confidence interval for $s$ and the most probable value. If ordered it can also reconstruct the significance. Technically, the analysis program can and usually is called more than once for each division, each time with different purpose. In most of its calls the analysis program does not know the true $\vec{a}$ and $\vec{b}$, when they are assumed to be unknown. However, in a special call, in order to compute the “precise” estimate of significance, for which the pseudo-experiments should be generated with the precise nuisance parameters, the main program additionally supplies the analysis program with the precise distributions of $\vec{a}$ and $\vec{b}$.

After the end of the loop by divisions for each “event” (that is for each main and auxiliary experiments) the main program can compare the intervals obtained for each division and choose the best according to any criteria. After the end of the loop by experiments it can check the coverage of these intervals and other characteristics. If the significance is studied, the main program can compare the approximate and the precise significances and allows one to make corresponding conclusions. It will be explained in the next sections that if the former if systematically better than the latter, the method is incorrect.

Most calculations for this work were made with $t_a = 0.25$, $N_a = 100$, $t_b = 5$, $N_b = 50$, $s_{\text{true}} = 2$. The behavior of almost all methods was also tested at microscopic dependency on signal in the conditions of $t_a = 0.25/20$.

The cases for 3, 5, 10 and 30 channels are shown in Fig. [1]. One can see that probability of observing zeros in the expected background is very high for many-channel cases. For just the last channel the probability of observing no events in it is around 12% for 5-channel case, 40% for 10 channel case and 76% for 30 channel case.

Most of the tests of intervals were performed in this paper for the one-sided confidence level of 90%. In some cases, especially for the cases without uncertainties, which are typically calculated very fast, the levels of 99% and even 99.845% (3σ level) were tested.

\[2\] Different values of these multiple parameters were not still scanned consistently with the final software versions of all methods because it is extremely time-consuming. But some partial tests with different parameters did not indicate that the conclusions might depend on them. Anyway, it is better to use methods that work at least at particular conditions, than to use methods that may not work at all.
2.3 Optimization of division

The most basic method of optimization is finding the division which provides the minimal interval width. This should usually be employed with some additional conditions, such as the absence of zeros in the expected-signal distribution, or another condition, depending on the method. The condition such as the former effectively excludes too many-channel divisions from consideration. The optimization by separate limits usually provides the shortage of coverage for any method. If the coverage for fixed divisions is 90%, usually something like 80% is obtained by this optimization.

In the case with nuisance parameter uncertainties the minimization of the interval width usually reduces the coverage with respect to the minimal coverage obtained at fixed divisions. Thus, if there is no noticeable margin in the latter, the coverage of the intervals optimized by widths can be slightly smaller than required. In the test example of this work the upper limit was susceptible for this problem, while the lower limit was usually correct for those methods that provide the correct lower limits for fixed divisions.

We will not consider here various other ways of optimization, that can eventually mix a channel with true zero expected background with the neighboring non-zero channel, thus potentially reducing the sensitivity of the experiment.

A simple way to provide the claimed optimized coverage is to request better fixed-division coverage. Then it needs to know for how much it should be better. In general this is an unclear issue.

A better way of improving the optimized coverage by the interval width of any non-Bayesian method is the use of the Bayesian credible intervals for obtaining the optimal division. Optimization by a different interval finding method, in a simplistic explanation, chooses the interval which is not always the shortest for given method, thus improving the coverage of given method. This optimization should not necessarily mix the zero-background channel with others. This method is used throughout this work.

If for particular conditions the lower limit is close to zero or almost always zero, it can be non-informative for the purpose of division optimization. In particular, this can happen if the confidence requested is very large, for example 99.9% for the test example studied in this work. The optimization by width is then reduced to the optimization by the upper limit and one can predict the lack of coverage as if optimized by the single upper limit. If the problem is caused by the extremely high confidence level, the optimization can be performed by the intervals obtained with a smaller level. Otherwise it is assumed, though not tested, that another criterion that indirectly indicates the distribution widths should be employed.

3 The Bayesian approach

3.1 The Bayesian probability density

There is a lot of discussions in the literature devoted to introduction to the Bayesian approach. For the purpose of this paper let us formulate it in the following way. Let us assume that the unknown parameter $s$ is a random value with the distribution $\pi(s)$, and it is unknown which particular value occurs at the time or the place of the experiment. Let us assume that the probability of observable $\bar{n}$ depends on $s$ and can be written as $P(\bar{n}|s)$. Then, the well known relations $p(s,\bar{n}) = p(s|\bar{n})p(\bar{n}) = P(\bar{n}|s)\pi(s)$ and $P(\bar{n}) = \int P(\bar{n}|s)\pi(s)\,ds$ indicate that if there is a set of experiments with $s$ distributed according to $\pi(s)$, the subset with $\bar{n}$ obtained has $s$ distributed according to

$$p(s|\bar{n}) = \frac{P(\bar{n}|s)\pi(s)}{\int P(\bar{n}|s)\pi(s)\,ds}. \quad (7)$$

This formula, and similar formula in discrete notations, is traditionally referred to as “the Bayes’ theorem” (See §8.7 in Ref. [24] and Ref. [10]).

The limits of integration here and later can be chosen either from $-\infty$ to $+\infty$ with $\pi(s) = 0$ at $s < 0$, or from 0 to $+\infty$.

In contrast, in the frequentist approach the parameter of interest is regarded as a constant from the viewpoint of the experimenter (see, for instance, p. 349 in Ref. [25]), but can vary arbitrary or even randomly from experiment to experiment (see p. 349 of the same source or §19.4 of Ref. [3]). The frequentist intervals cover the unknown true parameter with given probability in the sense that every interval obtained in each experiment should be compared with the actual value of the parameter of interest occurred in this particular experiment and should cover it with this probability. Despite of this, the frequentists do not make any statements about the probability of the unknown value (see §1.36 of Ref. [10]).

Footnote 3: We do not consider here extreme cases of the infinite intervals with 100% confidence level or of a priori empty intervals outside the working range. Such intervals are useless. Though, according to Ref. [26], all frequentist intervals are useless.
The frequentists argue by contradiction and reject what is (assumed to be) impossible. The Bayesians are looking for what is more or less possible from what is given. The Bayesian approach allows us to talk about the probability of unknown. The Bayesian probability for \( s \) to fall within a particular interval \([s_L, s_U]\) is given by \( \int_{s_L}^{s_U} p(s|\vec{n}) \, ds \).

### 3.2 The prior distribution for the parameter of interest

At the total absence of prior information we cannot give any preference to any specific value of \( s \). This idea is converted into the uniform or flat distribution \( \pi(s) \) (§8.19–8.20 of Ref. [24]) with an exception that in our case it has to be zero at \( s < 0 \). This solution is known to be not unique, if Eq. (7) can be rewritten as a function of some other variable \( r \) with a non-linear (non-unique in the discrete case) relation between \( s \) and \( r \). If \( \pi(r) \) is taken as uniform too, the result \( p(s) \, ds \) expressed in the terms of \( s \) will not be identical in general case to the result \( p(r(s)) \pi(r(s)) \, dr \) obtained through \( r \).

This ambiguity was a subject of long debate [16, 24, 27]. In order to obtain the identical result one has to use a non-uniform prior for \( r \) which assures that \( \pi(r) \) (which is constant in the given case) is proportional, according to the known “change-of-variables formula”, to \( \pi(r(s)) \pi(r(s)) \) (§5.35 in Ref. [16], §8.25 in Ref. [24]). For certain classes of transformations \( r(s) \), such as \( r = s^g \), where \( g \) is any non-zero power, one can find “invariant” priors that do not need to be changed to assure the constant results for these specific transformations. For the example of \( r = s^g \) this is \( 1/s^6 \). This creates an illusion that there is no need to select any specific parametrization.

However, there is no prior invariant in this sense for any possible transformation. Furthermore, there is no strong argument why the prior used in the analysis should be invariant at all. If a particular form of prior provides the same physical results for any parametrization, this does not make it special in any other sense except this. Indeed, the frequency interpretation of the Bayesian theorem described in the previous section implies the dependency on the prior anyway. The credible intervals depend on the prior too.

There are many other proposed alternative priors that depend on the shape of likelihood and on auxiliary parameters or their measurements. For example, “Jeffrey’s general rule” [27] leads to the so-called “reference priors”, which vary according to the change-of-variables formula, but depend on the shape of likelihood. This results in dependency of “prior knowledge” on particular experimental features, such as resolution, which is strange, at least. See more detailed analysis of this dependency in the introduction of Ref. [15] and also the discussion in Ref. [28] for arguments for and against.

The uniform prior is invariant at the transformations \( s = r + b \) with any \( b \) [27], but the condition \( s \geq 0 \) is neglected here. For given task there is neither need nor useful interpretation of any transformation of \( s \) to any other variable. The uniform prior does not shift the most probable value of \( s \) from the maximum likelihood value, thus preventing ambiguity, which of them is more “most probable”. Given the probability density for the uniform prior one can easily extract forecasts for any non-uniform prior by simple multiplication and renormalization.

Because of all these considerations the uniform prior was used for \( s \).

In the case of known \( \vec{a} \) and \( \vec{b} \) the value of \( P(\vec{n}|s) \), necessary for making calculations by Eq. (7), is simply expressed through \( P(\vec{n}|\vec{a}, \vec{b}, s) \).

### 3.3 The case of unknown \( \vec{a} \) and \( \vec{b} \)

If \( \vec{a} \) and \( \vec{b} \) are determined in an auxiliary experiment with finite precision, one can use their results \( \vec{n}_a \) and \( \vec{n}_b \) as the first approximation to \( \vec{a} \) and \( \vec{b} \). Then, instead of \( P(\vec{n}|s) \) in Eq. (7) one will use \( P(\vec{n}|\vec{a}, \vec{b}, s) \). Numerical checks have shown that for the considered example this approximation does not work.

The probability \( P(\vec{n}|s) \) can be more precisely expressed according to the complete probability formula by the convolution with probability densities of having certain values of parameters:

\[
P(\vec{n}|s) = \int \int P(\vec{n}, |s, \vec{a}, \vec{b}) \pi(\vec{a}|\vec{n}_a) \pi(\vec{b}|\vec{n}_b) \, d\vec{a} \, d\vec{b}. \tag{8}
\]

The probability densities of \( \vec{a} \) and \( \vec{b} \) can be reconstructed from the auxiliary measurements and expressed by the Bayes’ formula too:

\[
p(\theta_i|\theta_i) = \frac{P(n_{\theta_i}|\theta_i) \pi(\theta_i)}{\int P(n_{\theta_i}|\theta_i) \pi(\theta_i) \, d\theta_i}, \tag{9}
\]

where \( \theta \) stands for \( a \) or \( b \). The limits of integration are either from \( -\infty \) to \( +\infty \) with \( \pi(\theta_i) = 0 \) at \( \theta_i < 0 \), or from \( 0 \) to \( +\infty \).

Thus, the probability density \( p(s|\vec{n}) \) depends also on \( \vec{n}_a \) and \( \vec{n}_b \) and fully written as \( p(s|\vec{n}, \vec{n}_a, \vec{n}_b) \).

At substitution of Eqs. (9) into Eq. (8), and substitution of the result into Eq. (7) the denominators in Eqs. (9) are canceled and the result appears to be

\[
p(s|\vec{n}) = \frac{N(s)}{\int N(s) \, ds}, \tag{10}
\]

where

\[
N(s) = \int \int P(\vec{n}, |s, \vec{a}, \vec{b}) \pi(s) \times P(\vec{n}_a|\vec{a}) \pi(\vec{a}|\vec{n}_a) \pi(\vec{b}|\vec{n}_b) \, d\vec{a} \, d\vec{b}. \tag{11}
\]

Similar formula in different notations, with different nuisance parameters and initially, as a rule, with a non-factorized prior appears in many sources, see, for example, §3.5 of Ref. [10], Refs. [17, 18], and §1(b) of Ref. [25]. In
Table 1: Parameters of Bayesian posterior probability density distributions for the Poisson distribution of observations with different priors.

| Prior      | Mean | $\sigma^2$ | Maximum |
|------------|------|------------|---------|
| Uniform    | $n + 1$ | $n + 1$ | $n$ |
| $1/\sqrt{\mu}$ | $n + 0.5$ | $n + 0.5$ | $\max(n - 0.5, 0)$ |
| $1/\mu$     | $n$     | $n$       | $\max(n - 1, 0)$ |

our case the prior is automatically factorized. In addition, one can see that $P(\vec{n}_i|\vec{a})$ and $P(\vec{n}_b|\vec{b})$ are present in this formula together with $P(\vec{n}|s, \vec{a}, \vec{b})$. Interestingly, in this resulting formula there is no evident difference between the roles of the main and the auxiliary experiments, in contrast with their roles at its derivation.

On the other hand, the Eqs. (10, 11) show, at a careful investigation, that at the assumed presence of the background any main prior equal to any negative power of $s$ (thus infinite at $s = 0$) results in the infinite posterior probability density at $s = 0$. Forbidding these strange posteriors means forbidding such priors, which gives an additional argument for the use of the uniform main prior (c.f. §6.30 in Ref. [16]).

3.4 The prior distributions for the nuisance parameters

For the nuisance parameters $a_i$ and $b_i$ we need neither interval nor point estimates. Therefore some of the arguments from Section 3.2 are not applicable for choosing their priors $\pi(a_i)$ and $\pi(b_i)$. For this purpose Eq. (7) with substitution of Eqs. (8) and (9) or Eqs. (10, 11) give probability density of the true parameter. Obviously, the maximum of this function gives the most probable signal rate, which at $\pi(s) = \text{const}$ coincides with the maximum likelihood $P(\vec{n}|s)$ integrated over nuisance-parameters distributions. To calculate the most probable parameter of interest the researcher can compute the arithmetic average of the most probable signal rates calculated with uniform nuisance priors for the lower limit and with inverse nuisance priors for the upper limit. The average of this average over many pseudo-experiments appears to be very close to the true parameter.

3.5 Numerical calculations

When $1/\mu^p$, $0 < p < 1$ appears as a factor in an equation like Eq. (1), the multiplication of this prior with the Poisson distribution for zero observation diverges at $\mu \to 0$. To get useful result from Eq. (2) such prior is considered starting not from zero, but from a very small number, actually $10^{-100}$ for this paper, and is considered equal to zero below this limit. The results do not vary noticeably at varying the order of this power and hence do not depend on this specific choice. Similarly, all priors, inverse as well as uniform, are cut at some large enough value of $s$, above which all probability densities are effectively zero anyway.

Both numerical integration and analytic-based integration, in which the integrals were converted into some finite sums, were implemented. Precise and fast numerical integration of these functions with singularities appeared to be a complicated task. For the examples discussed the analytic-based integration works much faster.
one obtains $z$ for given p-value, which is the same as $\alpha$ in our notations, by $z = F^{-1}(1 - \alpha)$. If $s$ is the most probable value that maximizes $p(s|i\hat{n})$, see Eq. 10 or 11, the interval boundaries are set at the probability density smaller by the factor of $e^{-z^2/2}$. One has to find the lowest and the uppermost boundaries $s_L$ and $s_U$ such that $p(s_L|i\hat{n}) = p(s_U|i\hat{n}) = p(s|i\hat{n})e^{-z^2/2}$. If non-negative $s_L$ does not exist, it is equated to zero. These intervals appear to be close to the shortest intervals shown in Fig. 2 and have the same benefits and drawbacks.

The use of the right boundary of the central interval with the uniform prior and the left boundary of the central interval with the inverse prior provides coverage, but does not always provide the inclusion of the most probable value. The left boundary can never be precisely zero.

But all mentioned problems are solved if to take the right boundary from the central interval Eq. 12 computed with the uniform prior, and to set the left boundary at the same level of probability density as that for the right boundary, see Fig. 2. Graphically, one should draw a horizontal line from the upper edge of the right boundary to the left till its crossing with the distribution. If the non-negative $s_L$ satisfying this condition does not exist, it is equated to zero. If the left boundary of the classical central interval is lower for any reason, it has to be used instead of this modified boundary (for safety). Calculations of the simple one-channel problem without background for various $n$ show that this modified boundary will always be lower and usually almost equal to the lower boundary of the central interval for the inverse prior, which allows to conclude that it will not undercover. Hence these modified central intervals cover by both ends for this simple task.

In this method the probability of violating the lower limit can be smaller than $\alpha$. At small signal it can be even zero.

Unfortunately, despite all precautions such modified Bayesian central intervals do not guarantee the perfect coverage in the case of the uncertain expected signal and many channels with low or zero counts in the expected-signal histogram. But this is caused by other effects. It does not seem to be useful to correct it by manipulations with posterior probability densities.

When only the uncertainty of the expected background is present or dominates, this method provides coverage, at least in our test example. The next section will describe this in more detail.

3.7 The coverage, width and optimization of divisions for modified central intervals

If the nuisance parameters are known, the coverage of the Bayesian modified central intervals is provided for all fixed divisions, see Fig. 3.

At this figure and at the several next ones the $x$-axis is identical in all three plots and represents the number of channels (except the $x$-values at which the tails are plot-
The interval boundaries results in coverage about 83%. The optimization by separate uncertainties. The horizontal straight lines are results of optimization by interval widths. The optimization by separate interval boundaries results in coverage about 83%.

The points connected by the solid (dashed) line at the upper plot represent the coverage of the upper (lower) interval boundary.

At the middle plot the upper and lower limits are shown by the points connected by solid lines. The point estimates (the maxima of the Bayesian posterior) are the points connected by the dashed line. Obviously, the latter reside between the former.

To express that the optimized-division results are not linked to a certain fixed number of channels, they are plotted as horizontal straight lines, rather than the points which indicate the specific division.

The error tails at the uppermost coverage plots indicate the uncertainty of calculations for the standard 68% confidence level. These are frequentist uncertainties for binomial distribution of given outputs at the given number of experiments [30]. These uncertainties appear owing to the limited statistics of Monte Carlo simulations performed for this paper. At this particular plot they are very small due to very large statistics simulated. The Bayesian analysis without nuisance uncertainties is very quick, which makes it possible to calculate very many pseudo-experiments.

All errors plotted at two lower plots express the fluctuations of the respective values occurred experiment by experiment. To make the image more clear, the tails corresponding to upper limits are slightly inclined to the left, and the tails corresponding to lower limits to the right. The same inclination (not seen clearly in Fig. 3 because of too short tails) is present also at the coverage plots. The tails of values which are not associated with the specific boundary, that is the tails of point estimates and widths, are precisely vertical. The errors at two lower plots are calculated as the mean square deviations and hence correspond to standard 68% too.

The errors at the straight lines related to optimization are shown as the tails at some arbitrary chosen points at the right ends of these lines beyond the working 30-channel range.

Interestingly enough, the coverage presented at the uppermost plot, is almost constant and stays nearly 91% for all divisions for the studied example. But the interval widths and boundaries reach the plateau starting from 2 channels. For this case without nuisance uncertainties all the other reasonable methods behave similarly[4]. According to the similar calculations the lower boundary of the Bayesian central intervals (not modified) does not provide the stated coverage, as expected.

Thus, for the case with known nuisance parameters the modified Bayesian central intervals provide the frequentist coverage.

The mean point estimates in the middle plot almost coincide with the true value of the parameter of interest (2) both for fixed and for optimized divisions, so one cannot distinguish visually two dashed lines in this plot.

If the background uncertainties are switched on, the Bayesian method behaves as shown in Fig. 3. The upper and lower limits form a valley with narrowing in the middle. In the middle plot one can also see additional dotted lines, which display the mean point estimates (maximums of the posteriors) calculated with the safe priors appropriate for the upper and lower limit. It is seen that they are deviated from the optimal position simultaneously with the corresponding limits at the increase of the number of channels. Obviously this divergence is entirely due to the safe priors. Their arithmetic average drawn by the dashed line for fixed divisions and by the straight dashed line for optimized divisions is very close to the true s.

To save the place pictures with other priors are not shown, but looking at this figure one can imagine how it would look with the exchanged priors. The point estimates will be exchanged and each limit will follow the trend of its point estimate. The upper limit will fall and the lower limit rise. There will not be horizontal segments between 3

---

4 One can wonder, whether there exists a fundamental law which determines this kind of behavior.
and 5 channels as in Fig. 4. Correspondingly the coverage of both limits will be lost at 5 channels. At the medium plot the lower and the upper limit lines will cross each other at approximately 15 channels!

The priors $1/\sqrt{\mu}$ produce almost precise point estimates and not diverging limits for this example, but at the other parameters and at the greater number of channels they lead to deviations anyway. One might try to fit the power of $\mu$, but this is difficult to do knowing only the information from a single experiment. In this paper we do not pursue this purpose.

It is interesting that the frequentist (hybrid) methods that use safe priors behave qualitatively by the same way. As shown in Fig. 4 the optimization of the division by the interval width provides almost perfect 90% coverage for the Bayesian case. Obviously, the algorithm will usually take one of the medium divisions, which provides smaller intervals for given main and auxiliary experiment.

Thus for the case of unknown expected background the modified Bayesian central intervals with safe nuisance priors provide frequentist coverage, which is sometimes conservative.

When only the uncertainty of the expected signal is present, only the lower limit was found to cover the true parameter with no less than stated probability for all fixed divisions for this method, see Fig. 5 The coverage of the

upper limit falls from 92% for 1 channel to 86% for 30 channels. The same effect appears in all other methods that use the safe priors. The exchanged priors provide even worse coverage for the upper limit.

The reason of this pathology is simpler to illustrate for the Bayesian case. It has similar reasons for the other cases. The channels having the downward fluctuation of the expected signal and $n_{ai} = 0$ do not influence the result because in such channels $a_i$ is distributed very close to zero due to the use of the inverse prior and $f_i = t_a a_i s + t_b b_i$ does not depend on $s$. Only the rest of the channels, where $n_{ai}$ could fluctuate upward, influence the result. Since $a_i$ seems to be greater for such channels than it is in the average, smaller signal is enough to describe the observed result. Data indicate that not only zero $n_{ai}$ but also low non-zero values of $n_{ai}$ affect the result too. Apparently, the channels with downward fluctuations of $n_{ai}$ are stronger masked by the background and less participate in the result, than the channels with larger $n_{ai}$ do.

Choosing the division with the least width without zeros in the distribution of the expected signal allows one to obtain the upper limit with coverage slightly smaller than requested, as indicated by the horizontal solid line at the upper plot in Fig. 5. Some small lack of the coverage is deemed to be tolerable.

The fall of the average upper limit, as seen at the
medium plot, is not significant, from $\approx 2.8$ for 3 channels to $\approx 2.6$ for 30 channels.

One more argument for this method is that it seems to be little probable that one will ever have zero or close to zero content in a channel of the expected-signal distribution. This problem is more probable for the expected background.

When both uncertainties of expected background and signal are switched on, the behavior of all characteristics for the test example studied is qualitatively the same as for the case with the background uncertainty only. In particular, the upper limit covers for all fixed divisions. Obviously, at this set of parameters the uncertainty of the expected background and its treatment with the safe priors affects the results stronger than the uncertainty of the expected signal and its similar treatment does.

4 Frequentist Treatment of Maximum Likelihood Estimate

4.1 Introduction, the case without uncertainties.

Ciampolillo [19] and, independently, Mandelkern and Schultz [5] recently pointed out that the maximum likelihood estimate of the parameter of interest is a good test statistic for constructing the frequentist confidence intervals for the Poisson measurements with known expected signal and background. As they found, this test statistic allows one to avoid unphysical empty or nearly empty intervals at downward background fluctuations, from which the frequentist analyses with other statistics suffer.

In my opinion, it is very convenient to work directly with the estimates of the parameter of interest and its probability densities, in contrast with, for instance, various parameters of the methods based on likelihood ratios, whose meaning is not easy to understand and whose behavior is difficult to predict in practical situations. In these methods one should, as a rule, deal with the maximum likelihood estimates as intermediate values anyway. At some point it is natural to ask: why this estimate cannot be used as the test statistic itself? And it is immediately revealed that it can. In spite of this, this useful and intuitively convincing method attracted surprisingly little attention in the literature.

Here we will call this method “FML”, which means “Frequency of Maximum Likelihood”. The typical confidence belt for FML is shown in Fig. 6. This figure depicts the case of 5 channels with precisely known expected signal and background. The notation $\hat{s}$ means the value of $s$ that maximizes $P(\bar{n}|s)$, which is here equal to $P(\bar{n}|\bar{\theta}_{a}a_{s} + t_{b}) = \prod_{i=1}^{k}\text{Poisson}(n_{i}, t_{a}a_{s} + t_{b})$. It is assumed that $\bar{\theta}$ is chosen from non-negative values $[0, \infty]$. For each assumed or possible $s$ we can simulate a set of pseudo-experiments and obtain the distribution of $\hat{s}$, which is shown in this figure by the boxes with variable size. Each horizontal row of boxes depicts a histogram for a certain $s$. Technically, one can do this for each possible value of $s$ with some fine-grained mesh or one can prefer some more selective trials in order to reduce computer time expense.

After choosing a specific value of $s$ for the current trial one generates a set of pseudo-experiments with it. This process will be called subgeneration, in order to distinguish it from the generation of “real” experiments. In this work the latter are simulated by the method of Monte Carlo too, but this is done in the separated main program with the true parameters. In order to test the coverage the main program performs generation of “real” experiments with the true parameters. In order to determine the confidence interval for each “real” (generated) experiment by a frequentist method, the analysis program performs subgeneration of experiments for each tested $s$ with either known or assumed nuisance parameters, if they are unknown. In this section they are assumed to be known. For each “subgenerated” experiment the analysis program has to obtain $\hat{s}$ and fill it in the histogram.

The probability density distribution of obtaining $\hat{s}$ at given $s$ in the subgenerated experiment is denoted by $p(\hat{s} | s)$. The index $\gamma$ is included in order to indicate that the result
is obtained by subgeneration. The integrals
\[
\int_{\tilde{s}_{\text{right}}}^{\infty} p(\tilde{s}_i|s) \, d\tilde{s}_i = \alpha \tag{14}
\]
and
\[
\int_{0}^{\tilde{s}_{\text{left}}} p(\tilde{s}_i|s) \, d\tilde{s}_i = \alpha \tag{15}
\]
allow us to plot the boundaries of the confidence region, which are depicted in Fig. 6 by thick inclined solid trajectories $[U_0, U_4]$ and $[L_0, L_4]$.

For the measurement $C_2$ the confidence interval is given by $[L_2, U_2]$, which includes the true $s$ if it is depicted, for example, by $[A, B]$. For the measurement $C_3$ the confidence interval $[L_3, U_3]$ does not include the true $s$. The proof of the one-sided coverage of this lower limit is based on the idea that the probability for $L_3$ to be higher than $s$ is equal to the probability for $E_3$ to be to the right from $F_3$, and the latter is equal to $\alpha$ according to Eq. (14) or possibly smaller than $\alpha$ in the discrete case. Similarly one can prove the coverage on the upper limit, observing that the probability for $E_1$ to be to the left from $G$ is equal to $\alpha$ according to Eq. (15) or possibly smaller than that in the discrete case.

These trajectories look like lines and this was assumed in the paragraph above, but in reality they consist of small steps due to discreteness of the task. Let us search the solution of Eq. (14) by replacing $\tilde{s}_{\text{right}}$ by $\tilde{s}$ obtained for observed $\bar{n}$, which we denote here by $\tilde{s}_{\text{obs}}$. Then we have to fit $s$ to obtain the equality in this expression. Let us assume that the statistics of subgeneration is (nearly) infinite. Then the corresponding equation in the discrete form is
\[
\sum_{\tilde{s}_{\text{obs}} | \tilde{s}_{\text{obs}} \geq s} P(\tilde{n}_i|s, \tilde{a}, \tilde{b}) = \alpha, \tag{16}
\]
where $P(\tilde{n}_i|s, \tilde{a}, \tilde{b}) = \prod_{i=1}^{k} \text{Poisson}(n_{\gamma_i} t_{\alpha_i} s + t_{b_i})$. It is meant here that only those $\tilde{n}_i$ are included for which the condition under the sign of sum is satisfied. The solution $s$ of this equation is the lower confidence limit $s_L$. Note that this summing up should start from $\tilde{s}_{\text{obs}}$, not from the next allowed $\tilde{s}_i$ as might seem for the first glance. It is important because the value of this very sum calculated for $s = 0$ can be used as the significance of signal+background hypothesis versus the simple background hypothesis. Obviously, the greater $\tilde{s}_{\text{obs}}$ is, the more event is signal-like. Significance is traditionally estimated as the probability for the test statistic to exceed the observed value or to coincide with it.

If to neglect discrete effects, the precise coverage of intervals for any $\alpha$ is guaranteed by construction if and only if the subgenerated $\tilde{s}_i$ calculated at the true $s$ is distributed precisely as the experimental $\tilde{s}$ at the imaginary repetition of experiments. Since the true $s$ is unknown, this should also hold for any imaginary value that the true $s$ could take. The significance calculated by Eq. (16) is correct if this is hold for $s = 0$. This assumes that $\tilde{a}$ and $\tilde{b}$, which occurred in the experiment, are the same as used for subgeneration, that is they are known precisely.

Since $\tilde{s}$ is restricted to be non-negative, in the experiments where the formal $\tilde{s}$ chosen from $]-\infty, +\infty[$ is negative, $\tilde{s}$ chosen from $[0, +\infty[$ is usually zero. This results in appearance of a spike at zero in the distribution of $\tilde{s}$, which is described by the $\delta$-function with a certain weight. This weight is negligible at high $s$. At the reduction of $s$ this weight becomes larger and at some point greater than $\alpha$. In Fig. 6 this is crossing of the upper limit $[U_0, U_4]$ with the $y$-axis, that is the point $U_0$. There is no other choice but to exclude this spike entirely from the integral in Eq. (15) and to equate this integral to zero. Therefore the sign “=” in Eq. (15) has to be replaced by “<”.

If we want to keep constant the area inside the twosided interval, we have to shift the right boundary in order to cut off the right tail with 2$\alpha$ area instead of single $\alpha$. Thus the full lower boundary will pass through the points $L_5, L_6, L_7, L_1, L_2, L_3, L_4$. This very case was considered in [5, 19]. In this case the probability for the obtained lower limit to be higher than the true $s$ is unknown. It varied as a function of the true $s$ and can be either $\alpha$ or 2$\alpha$, depending on the position of the point $U_0$. If the latter is not determined and reported, it will be unknown too.

For comparison, boundaries of the shortest intervals and the low boundary of the modified central intervals in the Bayesian analysis have variable coverage too. But the coverage of them is directly calculated. Here the coverage cannot be directly calculated, and cannot be calculated at all, if one strictly follows the frequentist approach and does not consider the probability distribution of the true parameter of interest.

On the other hand, if the researcher does not shift the lower boundary when $s$ is below $U_0$, the coverage by the lower limit will be constant, but the simultaneous coverage of the true $s$ by both limits will be either $1 - \alpha$ or $1 - 2\alpha$. However, this two-sided coverage is less important in practice. There are exceptions, but usually this probability does not have any useful meaning. The violation of the lower and upper bound usually leads to different physical conclusions and their separate confidence levels are the values which are only important.

Thus, the confidence belt restricted by $[U_0, U_4]$ from above and $[L_0, L_4]$ from below will be tested in this research.

Calculations indicate that such technique provides plots almost identical to the plots obtained by the modified central Bayesian intervals for the case of known nuisance parameters, see Fig. 4. The one-sided coverage of both upper and lower boundaries for fixed divisions, as well as for the divisions optimized by the interval width, stays near 90% in all cases. Differences in the lower two plots are negligible. For the case with nuisance parameter uncertainties the method is split into many modifications, which will be described in the next section.
The significance is calculated directly for the case of known nuisance parameters. The case of unknown nuisance parameters requires more detailed consideration and is discussed in section 4.2.

4.2 The case of unknown \( \tilde{a} \) and \( \tilde{b} \)

If the values \( \tilde{a} \) and \( \tilde{b} \) are not known precisely, we have to use some approximations in the form of their assumed point values or probability densities. Similarly to the Bayesian case, the naive ignoring of these uncertainties and the use of \( \tilde{n} \tilde{a} \) instead of \( \tilde{a} \) and \( \tilde{n} \tilde{b} \) instead of \( \tilde{b} \) does not work well enough for FML in the example studied. More advanced assumptions or approximations are necessary.

They are needed at the maximum finding for the data of the real experiment (\( \hat{s} \)), at the subgeneration of the experiment and at the maximum finding by the ”subgenerated” data (\( \hat{s}_s \)). We will consider only the methods in which \( \hat{s} \) and \( \hat{s}_s \) are found in an identical way, because this will provide a feature that will be later called “modeled coverage”. It differs from the frequentist coverage by assuming some (probably fictitious) values or distributions of the nuisance parameters at the subgeneration and possibly at the calculation of \( \hat{s} \) and \( \hat{s}_s \) and/or selection of experiments with fixed auxiliary outcomes \( \tilde{n} \tilde{a} \) and \( \tilde{n} \tilde{b} \). Whether the frequentist coverage is provided, expected, assumed, or not assumed at all, this modeled coverage is an important feature itself. Arguably, it can be considered as sufficient, if the frequentist coverage is unknown, but the model for the nuisance parameters is reasonable.

Nevertheless, there are very many possible variants of these approximations. For instance, one can obtain the optimal nuisance parameters that best describe the real data, namely those which maximize \( P(\tilde{n}, \tilde{n}_a, \tilde{n}_b | \tilde{a}, \tilde{b}) \), that is \( \max_{\tilde{a}, \tilde{b}} \) \( P(\tilde{n}, \tilde{n}_a, \tilde{n}_b | \tilde{a}, \tilde{b}) \). The global maximization is proposed in Ref. 19 (p. 1421) and in more recent Ref. 31. Here \( s \) is floating and fitted too. Let us denote these values \( \hat{s}_{\text{obs}}, \tilde{a}_{\text{obs}} \) and \( \tilde{b}_{\text{obs}} \). The value \( \hat{s}_{\text{obs}} \) has to be used as the observed test statistic value. The subgeneration of the main experiment can be performed with these point values \( \hat{s}_{\text{obs}}, \tilde{a}_{\text{obs}} \) and \( \tilde{b}_{\text{obs}} \) by substituting them into Eq. 2 and using the result as the parameter of the Poisson distribution Eq. 5. According to which one has to generate random numbers \( \tilde{n}_s \). When analyzing subgenerated experiments one can use these values for finding \( \hat{s}_s \) by maximizing \( P(\tilde{n}, \tilde{n}_a, \tilde{n}_b | \tilde{a}, \tilde{b}) \) with fixed \( \hat{s}_{\text{obs}}, \tilde{a}_{\text{obs}}, \) and \( \tilde{b}_{\text{obs}} \), although these \( \hat{s}_{\text{obs}}, \tilde{a}_{\text{obs}}, \) and \( \tilde{b}_{\text{obs}} \) are not always optimal for the different subgenerated data. One can also search for optimal \( \hat{s}_s \) with floating \( \tilde{a}_{\text{obs}} \) and \( \tilde{b}_{\text{obs}} \), thus allowing them to differ from the most probable values observed for the real data. In this case one maximizes \( P(\tilde{n}, \tilde{n}_a, \tilde{n}_b | \tilde{a}, \tilde{b}) \) in Eq. 17, assuming that it is the maximum of Eq. 7 with replacement of \( P(\tilde{n}_s | \tilde{a}, \tilde{b}) \) by \( P(\tilde{n}_s | \tilde{a}, \tilde{b}) \). We know that \( n_i \) was subgenerated with certain \( \tilde{a} \) and \( \tilde{b} \), but we do not know with which \( a \) and \( b \) the value of \( n_i \) was “generated” in nature. So there is no analogous replacement for \( \hat{s}_{\text{obs}}(\tilde{n}, \tilde{n}_a, \tilde{n}_b) \).

The p-value determined by the left hand side of Eq. 17 at \( s = 0 \) belongs to the category of “prior predictive p-values” 33, 34.
The limit $s = s_L$ that satisfies this equation will depend on the measured $\vec{n}_a$ and $\vec{n}_b$. For the upper limit we will have one value of $s_{\text{obs}}$ found with the safe priors valid for the upper limit, and for the lower limit we will have another value of $s_{\text{obs}}$ found with different safe priors valid for the lower limit. Then for the upper limit we have to use one line $[C_i, D_i]$, see Fig. 6, and for the lower limit another $[C_i, D_i]$. Other variants of priors do not work satisfactorily.

Let us assume that the nuisance parameters are random in reality and are distributed according to $p(\vec{a}_i|\vec{n}_a)$ and $p(\vec{b}_i|\vec{n}_b)$ given by Eq. (9). So if the main experiment is repeated many times, the experimental $\vec{n}_j$ obtained at the j-th repetition should be caused by the unknown random $\vec{a}_j$ and $\vec{b}_j$ that are sampled from these distributions. It is assumed that the researcher knows these distributions and can use them for fitting of $s$ and $s_\gamma$. Let us assume that the auxiliary measurements are not repeated anymore. On the other hand, they can be repeated but only those main experiments are chosen for analysis, for which the auxiliary measurements $\vec{n}_a$ and $\vec{n}_b$ coincide with primary $\vec{n}_a$ and $\vec{n}_b$, respectively. We have right to do such selection. In both cases the distributions of $\vec{a}_j$ and $\vec{b}_j$ turn out to be identical with those used at the analysis of the primary observed data. This has two consequences. First, the distribution of $s_{\text{obs}}$ accumulated in a long range of such experiments will be identical to the distribution of $s_\gamma$ used at the analysis of the primary experiment. Second, the confidence belts that have to be used in the analysis of any such “secondary” experiments turn out to be identical to that used at the analysis of the primary experiment. The vertical projection of the identical $s$-distributions onto the sides of the confidence belt, $[U_0, U_4]$ and $[L_0, L_4]$ in Fig. 6, for instance, the projection of $E_i$ into $U_i$ and $L_i$), lead to identical distributions of the reconstructed limits and hence to identical coverage of the true parameter. Thus this model for the nuisance parameters provides the coverage by construction for this method of analysis. This feature can be understood as the interpretation of this method. The $p$-values obtained by this method are interpreted in the same way.

In general, assuming that the nuisance parameters fluctuate according to some posterior distribution is a usual way of taking into account systematic errors of all sorts, although an assumption of random nuisance parameters is sometimes not realistic.

Thus, this method provides a good modeled coverage, but whether it yields the frequentist coverage has to be shown in numerical calculations.

Calculations have shown that this method has characteristics that are very similar to those of the Bayesian modified central intervals, described earlier. For example, the case with uncertainty of the expected background is shown in Fig. 7. The optimization with the Bayesian modified central intervals provides reasonably good coverage of the optimized limits. The optimization with the intervals obtained by this method provides slightly worse coverage of the upper limit. This method, as well as the Bayesian one, produces two point estimates, which has to be averaged.

This variant of FML can be briefly denoted by FMML, “Frequency of Marginalized Maximum Likelihood”, or more explicitly by SSP–FMML, where the prefix means the Sub-generation with Safe Priors. Other priors do not work satisfactory. For example, some figure in the following will show the case with Sub-generation with Exchanged Priors, or SEP–FMML (uniform for upper limit and inverse for lower one).

Another alternative approach to (SSP–)FMML is finding the global maximum of the common likelihood given by Eq. (11) and expressed by $P(\vec{n}_i, \vec{n}_a, \vec{n}_b|s_{\text{obs}}, \vec{a}_{\text{obs}}, \vec{b}_{\text{obs}})$ for the case of the observed data, instead of the maximum of the Bayesian posterior as it is required for FMML. The subgeneration is performed precisely as for the SSP–FMML. At the analysis of the subgenerated experiments it needs to find the global maximum $P(\vec{n}_i\gamma, \vec{n}_a, \vec{n}_b|s_{\text{obs}}, \vec{a}_{\text{obs}}, \vec{b}_{\text{obs}})$ with floating $s_{\gamma}, \vec{a}_\gamma, \vec{b}_\gamma$. Eq. (17) is not changed, except that the values $s_{\gamma}(\vec{n}_i, \vec{n}_a, \vec{n}_b)$ and $s_{\text{obs}}(\vec{n}, \vec{n}_a, \vec{n}_b)$ have different sense, which is described above. In this method the Bayesian priors are used only for the subgeneration. At the maximization the priors are not used. Hence the observed $s_{\text{obs}}(\vec{n}, \vec{n}_a, \vec{n}_b)$ is single and should not be averaged to obtain the final point estimate as necessary for the Bayesian and FMML cases. It has some systematic shift,
but the latter is not large.

This variant of FML can be called SSP-FGML, Subgeneration with Safe Priors, Frequency of Global Maximum Likelihood.

The interpretation of this method is similar to that for FMML. It provides the coverage if the nuisance parameters were random in reality and were distributed according to the results obtained with Eq. (2). Obviously, the method of finding \( \hat{s} \) is different compared with that of FMML.

Calculations indicate that all performance characteristics of SSP–FMML and SSP–FGML are almost the same with three exceptions which are worth mentioning. First, the upper and the lower limit diverge less at 30 channels for SSP–FGML than they do for SSP–FMML, see Fig. 7. Instead of the average interval width equal to approximately 3.8 units for SSP–FMML, SSP–FGML gives about 3.4 units. Second, SSP–FGML is more stable for incorrectly set nuisance priors. Third, it is faster in calculations by the currently existing software. However, SSP–FGML finds the most probable values of the main parameter taking into account the most probable nuisance parameters and ignoring the other possible values, whereas SSP–FMML takes into account all chances for them. The latter is more appealing both conceptually and also technically. One can envisage that in some cases the nuisance parameter can be predicted from general theoretical considerations as an interval of allowed values, assuming equal probabilities inside this range and zero outside it. Then, in the case of SSP–FGML there may be no single most probable values \( \hat{\alpha}_{\text{obs}} \) and \( \hat{\beta}_{\text{obs}} \), and, worse, no single \( \hat{s}_{\gamma} \). The most probable \( \hat{s} \) can be represented by the intervals of equally probable values. It is then unclear how to find the confidence interval for it. This problem seems to be impossible for SSP–FMML.

Note also that in the case of such non-measured, but theoretically predicted nuisance-parameter distribution of any shape, in both FMML and FGML, the researcher should simply take this distribution as is, without the priors.

In both FMML and FGML the auxiliary measurements are not generated at the subgeneration stage. As indicated above, the subgeneration of the main measurements in both methods and fitting in FMML is performed with the assumed distribution of nuisance parameters. The fitting in FGML is performed with really observed auxiliary measurements. The question is whether one could obtain a method with better characteristics which uses the “subgenerated” auxiliary measurements. No such method was found, but for the sake of completeness some not too bad methods are described below.

Let us consider the following algorithm. The most probable nuisance parameters \( \hat{\alpha}_{\text{obs}} \) and \( \hat{\beta}_{\text{obs}} \) are determined by the global maximization of \( P(\hat{n}, \hat{n}_{\alpha}, \hat{n}_{\beta} | \hat{s}_{\alpha_{\text{obs}}, \hat{\beta}_{\text{obs}}} ) \) (Eq. (5)) for generation of random \( \hat{n}_{\alpha_{\gamma}} \) and \( \hat{n}_{\beta_{\gamma}} \). The subgeneration of the main experiment is performed by substituting \( \hat{\alpha}_{\text{obs}} \) and \( \hat{\beta}_{\text{obs}} \) into Eq. (2) and using the result as the parameter of the Poisson distribution, according to which one has to generate \( \hat{n}_{\gamma} \). Then the ordinary FGML is applied to these data. For the subgenerated experiments it needs to find the global maximum \( P(\hat{n}_{\gamma}, \hat{n}_{\alpha_{\gamma}}, \hat{n}_{\beta_{\gamma}} | \hat{s}_{\gamma}, \hat{\alpha}_{\gamma}, \hat{\beta}_{\gamma} ) \) with floating \( \hat{s}_{\gamma}, \hat{\alpha}_{\gamma}, \hat{\beta}_{\gamma} \).

This method will be called here SRN–FGML, where the prefix means Subgeneration with Random Nuisance.

This method does not provide the modeled coverage. Indeed, if the main experiment is repeated, \( \hat{\alpha}_{\text{obs}} \) and \( \hat{\beta}_{\text{obs}} \), which are used for subgeneration after each repetition, will be different whether one repeats the auxiliary experiments or not. This means that the confidence belt will be different each next time. Let us imagine that the true values of \( a \) and \( b \) are equal to \( \hat{\alpha}_{\text{obs}} \) and \( \hat{\beta}_{\text{obs}} \) obtained at the analysis of initial or real experiment and the auxiliary experiments are repeated too. Then the distribution of \( \hat{s} \) obtained in such imaginary repetitions will be identical to that of \( \hat{s}_{\gamma} \) used in the analysis of the primary experiment. However, each next j-th imaginary experiment will produce different \( \hat{n}_{\gamma}, \hat{n}_{\alpha_{\gamma}}, \) and \( \hat{n}_{\beta_{\gamma}} \), which will yield different most probable \( \hat{\alpha}_{\text{obs}_{\gamma}},j \) and \( \hat{\beta}_{\text{obs}_{\gamma}},j \), which will result in different sequences of \( \hat{n}_{\gamma}, \hat{n}_{\alpha_{\gamma}}, \) and \( \hat{n}_{\beta_{\gamma}} \), which will lead to different confidence belts. Thus the modeled coverage is not provided by this method and, hence, the interpretation of its results is not clear.

Calculations indicate that the lower limit calculated by this method does not provide the frequentist coverage too. In addition the method supplies too optimistic significances, which, if averaged over a certain sequences of experiments, appear to be better than the average “precise” significance, see the next section for the details.

The method that gives better frequentist coverage refits \( \hat{\alpha}_{\text{obs}} \) and \( \hat{\beta}_{\text{obs}} \) by finding their values, now denoted by \( \hat{\alpha}_{\text{obs}} \) and \( \hat{\beta}_{\text{obs}} \), that maximize \( P(\hat{n}, \hat{n}_{\alpha}, \hat{n}_{\beta} | \hat{s}, \hat{\alpha}_{\text{obs}}, \hat{\beta}_{\text{obs}} ) \) for each given \( s \). Similar approach is used in the LHC-style CLs method, which will be described later. The values \( \hat{\alpha}_{\text{obs}} \) and \( \hat{\beta}_{\text{obs}} \) can be looked as adjusted for given \( s \), which changes the abbreviation from SRN to SARN: Subgeneration with Adjusted Random Nuisance. These values are used for subgeneration of both the main experiment and the auxiliary experiments. Then ordinary FGML is used for finding \( \hat{s} \). Thus, \( \hat{\alpha}_{\text{obs}} \) that maximize \( P(\hat{n}, \hat{n}_{\alpha}, \hat{n}_{\beta} | \hat{s}_{\alpha_{\text{obs}}, \hat{\beta}_{\text{obs}}} ) \) is used as the threshold and \( \hat{s}_{\gamma} \) that maximize \( P(\hat{n}_{\gamma}, \hat{n}_{\alpha_{\gamma}}, \hat{n}_{\beta_{\gamma}} | \hat{s}_{\gamma}, \hat{\alpha}_{\gamma}, \hat{\beta}_{\gamma} ) \) is used for comparison. Eq. (16) can be rewritten as

\[
\sum_{\hat{n}_{\gamma}, \hat{n}_{\alpha_{\gamma}}, \hat{n}_{\beta_{\gamma}}} \left[ P(\hat{n}_{\gamma} | \hat{s}_{\gamma}, \hat{\alpha}_{\gamma}, \hat{\beta}_{\gamma} ) \times P(\hat{n}_{\alpha_{\gamma}} | \hat{\alpha}_{\gamma} ) P(\hat{n}_{\beta_{\gamma}} | \hat{\beta}_{\gamma} ) \right] = \alpha. \tag{18}
\]

If \( s \) grows, \( \hat{\alpha}_{\text{obs}} \) and \( \hat{\beta}_{\text{obs}} \) have to decrease in general case, and vice versa. The variation of \( \hat{\alpha}_{\text{obs}} \) and \( \hat{\beta}_{\text{obs}} \) will lead
to corresponding increase or decrease of the average $\bar{n}_s$, $\bar{n}_a$, $\bar{n}_b$, and $\bar{n}_{\gamma}$. Since all data are simultaneously changed, it is rather difficult to deduce how this should shift the reconstructed limits. Without any numerical tests it is hard to tell which method is better, which method works or does not work.

What is worse, this method, as well as the previous one, does not provide the modeled coverage by construction. The distribution of $\delta_{\text{obs}}$ in a long range of experiments coincides with the distribution of $\delta_{\gamma}$ calculated at $s = s_{\text{true}}$ only if the true $\bar{a}$ and $\bar{b}$ coincide with $\hat{a}_{\text{obs}}$ and $\hat{b}_{\text{obs}}$ respectively, calculated at $s = s_{\text{true}}$ for the data of the real main and auxiliary experiments. However, the confidence belt fluctuates by the same reasons as for the SRN–FGML method and therefore the coverage is not provided automatically. The interpretation of the $p$-value is not affected by the varied confidence belt and it is provided. So the estimate of the $p$-value by this method is true provided that the true $\bar{a}$ and $\bar{b}$ coincide with $\hat{a}_{\text{obs}}$ and $\hat{b}_{\text{obs}}$ respectively at $s = 0$. But since this interpretation is based on single values of $\bar{a}$ and $\bar{b}$ without considering alternatives, it can scarcely be convincing.

The frequentist tests show the following features. In the case when only the uncertainty of the expected background is present, the upper limit of SARN–FGML behaves similarly to that of SSP–FMML and SSP–FGML from 1 to 3 channels, but at the further dividing of channels it does not exhibit steady growth by the average position and the coverage, in contrast with the behavior of SSP–FMML and SSP–FGML. But the lower limit losses the coverage when the number of channels increases. When only the uncertainty of the expected signal is present, the coverage is good, except for the last 30-channel case, where the coverage of the lower limit shows some slight reduction below the necessary level. So at the optimization it needs to forbid zeros in the expected background, not in the expected signal as for SSP–FMML, SSP–FGML and for the Bayesian analysis. Thus this method cannot be recommended as general, since it would give poor performance at the presence of the truly zero channels in the expected background distribution. It will be shown later that this method also gives slightly too optimistic estimates of significance at the presence of the expected-signal uncertainties only. Moreover, since this method is based on the global maximization, it may be difficult to include in it theoretical predictions of the nuisance parameters in the form of intervals, as well as for SSP–FGML. Therefore, it needs to conclude that this method cannot be recommended.

At the presence of both uncertainties it is possible to combine SSP–FMML or SSP–FGML treatment of the expected-background uncertainty with SARN–FGML treatment of the expected-signal uncertainty. If one of these uncertainties dominates, the method will be reduced to SSP–FMML (FGML) and SARN–FGML respectively and will have all drawbacks inherent to either of them. Because of the problems with significance and intervals inherent to SARN–FGML this combined method cannot be recommended too.

### 4.3 Numerical calculations

As well as in the Bayesian case, which is very close to FMML, both marginalization (for FMML) and the global maximization (for FGML) was found to be faster computed with elements of analytic calculations. For the case of maximization see, for instance, Ref. [13] for background-maximization formula. To reduce computational overhead the direct building of the confidence belts is reasonable to replace by some algorithm of fitting the limits for specific observations. Of course, for fixed nuisance parameters one can calculate the confidence belt once and forever, but this cannot be done for the unknown nuisance parameters. Many precautions should be taken to avoid various stochastic fluctuations and problems caused by discreteness. Description of details would exceed the reasonable size of this paper. Technically, the upgraded software package [35] was used for all fitting and equation-solving tasks mentioned in this paper. In order to avoid treating data samples that were already treated the analysis of each experiment is memorized and used at the treatment of the next experiments. A special program either extracts the ready solution or sets closer boundaries for faster fitting.

### 4.4 Significance for the case of unknown $\bar{a}$ and $\bar{b}$

It was already mentioned in section[1.2] that in the case of known nuisance parameters Eq. (16) at $s = 0$ gives the significance of signal versus background. In the case of unknown nuisance parameters this equation is replaced by Eqs. (17) or (18), but they are not precise since they use for subgeneration the assumed values or distributions of $\bar{a}$ and $\bar{b}$ instead of the precise values. It was also described, that not all, but rather a few methods allow the user to obtain intervals with coverage in the case of unknown nuisance parameters. The insufficient coverage of the lower limits means that the left hand sides of Eq. (17) or (18) are smaller than they should be in reality for given $\delta_{\text{obs}}$. The line $[L_0, L_1]$ calculated for certain $\alpha$ is shifted to the left and upward direction. Therefore the limit found by it covers the true $s$ less frequently than assumed. If the left hand sides of Eq. (17) or (18) are smaller than they should be in reality, the estimate of $p$-value by them should also be smaller, than necessary, too. This means that the estimate of significance should be too optimistic and hence incorrect. Thus, if the coverage of the lower limit is smaller than stated, the $p$-value is probably smaller than it should be too, that is it is probably too optimistic and not reliable.

Let us denote the “precise” $p$-value for SSP–FMML and SSP–FGML for given $\bar{n}$, $\bar{n}_a$ and $\bar{n}_b$ as $\rho_p(\bar{n}, \bar{n}_a, \bar{n}_b)$,
where the index “p” means “precise”:

\[ \rho_p(\vec{n}, \vec{n}_a, \vec{n}_b) = \sum_{\vec{n}_{\gamma p}: \tilde{s}_s(\vec{n}_{\gamma p}, \vec{n}_a, \vec{n}_b) \geq \tilde{s}_a(\vec{n}_a, \vec{n}_b)} P(\vec{n}_{\gamma p}|s = 0, \vec{u}, \vec{b}). \]

Here \( P(\vec{n}_{\gamma p}|s = 0, \vec{u}, \vec{b}) = \prod_{i=1}^k \text{Poisson}(n_{\gamma pi}, t_b b_i) \). This is equivalent of saying that \( \vec{n}_{\gamma p} \) is subgenerated with the true nuisance parameters \( \vec{b} \). In this sense \( \vec{n}_{\gamma p} \) is “precise” and therefore it is marked by the additional index “p” as well. The value \( \rho_p \) is precise only in this sense, because what means the precise \( p \)-value in general for the case of uncertain nuisance parameters is undefined. At the fitting of \( \tilde{s} \)-values the value \( \rho_p \) takes into account that the nuisance parameters are unknown. The value \( \vec{n} \) is also generated with the true parameters \( \vec{a} \) and \( \vec{b} \), but here \( \vec{n} \) is predefined, as well as \( \vec{n}_a \) and \( \vec{n}_b \). This value \( \rho_p \) should be obtained for given test statistic \( \tilde{s}_{ob}(\vec{n}, \vec{n}_a, \vec{n}_b) \) in the case of future replication of experiments with true parameters.

Similarly for the case of SRN–FGML and SARN–FGML one can write

\[ \rho_p(\vec{n}, \vec{n}_a, \vec{n}_b) = \sum_{\vec{n}_{\gamma p}: \tilde{s}_s(\vec{n}_{\gamma p}, \vec{n}_a, \vec{n}_b, \vec{n}_{\gamma bp}) \geq \tilde{s}_a(\vec{n}_a, \vec{n}_b, \vec{n}_{\gamma bp})} P(\vec{n}_{\gamma p}|s = 0, \vec{u}, \vec{b}) \times P(\vec{n}_{\gamma bp} | \vec{a}) P(\vec{n}_{\gamma bp} | \vec{b}). \]

The average “precise” \( p \)-value for given \( \vec{n} \) will be

\[ \overline{\rho}_p(\vec{n}) = \sum_{\vec{n}_a, \vec{n}_b} P(\vec{n}_a | \vec{n}) P(\vec{n}_b | \vec{n}) \rho_p(\vec{n}, \vec{n}_a, \vec{n}_b). \quad (19) \]

The average approximate \( p \)-value for given \( \vec{n} \) can be found similarly to the “precise” case:

\[ \overline{\rho}_a(\vec{n}) = \sum_{\vec{n}_a, \vec{n}_b} P(\vec{n}_a | \vec{n}) P(\vec{n}_b | \vec{n}) \rho_a(\vec{n}, \vec{n}_a, \vec{n}_b). \quad (20) \]

Here \( \rho_a(\vec{n}, \vec{n}_a, \vec{n}_b) \) is the left hand side of either (17) or (18) (or mix) at \( s = 0 \), depending on the method. It can be obtained in the real experiment, but the average cannot.

Similarly one can find average significances in z-units: “precise” \( z_p \) and approximate \( z_a \), if \( \rho_p(\vec{n}, \vec{n}_a, \vec{n}_b) \) and \( \rho_a(\vec{n}, \vec{n}_a, \vec{n}_b) \) are converted to corresponding \( z \) through known formula \( z = F^{-1}(1 - \rho) \), where \( F \) is defined in Eq. (13), and \( z \) are averaged according to Eq. (19) or (20). One can proceed and find the total averages of \( \rho \) and \( z \) over the true distribution of \( \vec{n} \) too. This should result in less sensitive tests.

If there exists \( \vec{n} \) for which \( \overline{\rho}_a(\vec{n}) < \overline{\rho}_p(\vec{n}) \) or \( \overline{\rho}_a(\vec{n}) > \overline{\rho}_p(\vec{n}) \), the method is doubtful. This means that the nuisance-parameter uncertainties spoil the determination of the significance for this particular main measurement. Similar conclusion for any main measurement can be drawn if the total averages are in these relations. The researcher usually deals with \( z \)-values, therefore the latter is perhaps more important.

If one sees that \( \rho \) or \( z \)-values are in these relations, he knows that this significance estimate (expressed in this or that way) is better in average than the significance occurring in nature for the subset of experiments with this \( \vec{n} \) or for the full set of experiments. Obviously, such approximate estimates should be declined as too optimistic.

Note that it is not recommended to compare particular values \( \rho_p(\vec{n}, \vec{n}_a, \vec{n}_b) \) and \( \rho_p(\vec{n}, \vec{n}_a, \vec{n}_b) \) or corresponding \( z \)-values, since the fluctuations of the nuisance parameters cannot be fully recovered at the calculation of particular \( \rho_p(\vec{n}, \vec{n}_a, \vec{n}_b) \) and it can apparently be smaller than \( \rho_a(\vec{n}, \vec{n}_a, \vec{n}_b) \) for any method of analysis.

In general it is very difficult to calculate the average \( p \) or \( z \)-values with good precision due to large computer time expenses. Only a very limited number of events were calculated in this work. Note that in order to reduce the statistical errors it is reasonable to calculate the differences \( \overline{\rho}_a(\vec{n}) - \overline{\rho}_p(\vec{n}) = \rho_a(\vec{n}) - \rho_p(\vec{n}) \) and \( \overline{\rho}_a(\vec{n}) - \overline{\rho}_p(\vec{n}) = z_a(\vec{n}) - z_p(\vec{n}) \), but not the individual averages. In our case it appeared to be possible to reduce the statistical uncertainty to the level of evidence (3σ) or sometimes discovery (5σ). If these differences have wrong sign with this significance, this gives the proof that the method is formally incorrect.

For SSP–FMML and SSP–FGML the experimental outputs \( \vec{n} \) that result in wrong signs of these expressions were not found. The total averages are good too, with very good margin. Thus these methods have correct significance.

The same methods with wrong priors, SEP–FMML and SEP–FGML, give bad results as expected from the insufficient coverage provided by their lower limits.

The pathological experimental outputs that provide small but negative \( z_p(\vec{n}) - z_a(\vec{n}) \) for the case of expected-signal uncertainty were found for SARN–FGML with statistical significance of calculation more than 5σ. It is strange that there is no pathology found for the coverage of intervals in this case, but the differences of \( z \) are small anyway. So the connection between the coverage of intervals and significance is not strict. If separate tests of intervals and significance are possible, they should both be performed.

Fig. 8 shows example of a main experiment \( \vec{n} \), repeated with a sequence of random auxiliary experiments and treated by different methods including those which will be described later. This is computed for the case of expected-background only uncertainty.

It displays the mean differences of \( z \)-values. It is seen that the good results are produced only by SSP–FMML, SSP–FGML, and SARN–FGML. Remarkably, only SSP–FMML and SSP–FGML provide the correct coverage at fixed divisions for the case of expected-background uncertainty. Here SARN–FGML is looking good too, but similar investigation with the expected-signal uncertainty declines it. For the latter case these differences are precisely zero for SSP–FMML and SSP–FGML, so these methods are always accepted when only the uncertainty of the expected signal is present.

In the expressions above the fixed division into channels was assumed. If at each next experiment one can choose different division, all vectors will have variable di-
In some references by \( \lambda \), background are switched on, the expected signal is known precisely. For SARN–MR–CL, only so they cannot be smaller than zero. In principle, all parameters are limited to their physical values. As mentioned earlier, the probability density, given, for instance, by Eq. (4), can be considered as the likelihood of the parameters. We will not use here an additional notation for it (usually \( L \)). The likelihood ratio is denoted in some references by \( \lambda \) and defined by

\[
\lambda(s) = \frac{P(\vec{n}, \vec{n}_a, \vec{n}_b|s, \vec{a}, \vec{b})}{P(\vec{n}, \vec{n}_a, \vec{n}_b|s, \vec{a}, \vec{b})}.
\]  

Here the denominator is maximized by all parameters, and the nominator is maximized by the nuisance parameters only for the specified \( s \). Here, as well as everywhere in this paper, all parameters are limited to their physical values only, so they cannot be smaller than zero. In principle, this method can be applied also without this restriction, as in Refs. [8, 10], but here this option is not considered as having unclear physical sense.

Given \( \alpha \) one should obtain \( z \) as described after Eq. (13). Then the lowest \( s_L \) and the uppermost \( s_U \) boundaries that satisfy \( \lambda(s_L) = \lambda(s_U) = e^{-z^2/2} \) are taken as limits. Alternatively, one can obtain the same limits from the fractile of the \( \chi^2 \)-distribution with one degree of freedom, as recommended in Ref. [10]. Given \( 2\alpha \) one obtains \( z^2 \) and proceeds by the same way. If non-negative \( s_L \) does not exist, it is equated to zero.

An attractive feature of this method is the absence of priors. There are many reports about the good coverage of this method for particular tasks. But formally the coverage is not guaranteed. The tests with the example studied here show that the coverage is not stable and sometimes not sufficient, see Fig. 9. Optimization makes it even worse. The “confidence” interval is shorter, but it remains of the same order of magnitude as the intervals obtained by the Bayesian method, SSP–FMML, and SSP–FGML, see Fig. 4. So this method can be used for fast rough estimates of intervals. Obviously, similarly to the Bayesian method, it cannot provide significance in a

Figure 8: The difference between “precise” and “approximate” \( z \)-values for 30 events with fixed arbitrarily chosen main measurement (the same for each plot) and random auxiliary measurements (the same sequence for each plot). For SARN–MR–CL, there are 300 simulated experiments (to obtain better precision). Only the uncertainties of the expected background are switched on, the expected signal is known precisely. For SARN–MR–CL, the 3-channel and 30-channel divisions have negative differences determined with statistical significances of 11.9 and 9.0 respectively (better than the usual 5\( \sigma \) discovery level).
The parameter one-sided “confidence” (α = 0.1), the uncertainty of the expected background, the known expected signal. The horizontal lines depict optimization by the width of these intervals.

direct way. Its frequentist reformulation, capable to do so, will be described in the next section.

6 Frequentist treatment of the likelihood ratio and the “CL_s method”

6.1 Introduction, the case without uncertainties

The Neyman-Pearson theorem (see Ref. [36] or brief proofs in Ref. [7] and §20.10 of Ref. [3]) says that the likelihood ratio is the best test statistic for distinction between two hypotheses. Following notations of Read [7] we now denote it by Q and write

\[ Q = \frac{P(\vec{n}|s)}{P(\vec{n}|s_{\text{ref}})}. \]  

(22)

Here we ignore so far the issues with nuisance parameters. The parameter s_{\text{ref}} is some reference value of s. According to the initially proposed approach \[ s_{\text{ref}} = 0. \] Here this approach will be called the classical or Background-Related method and denoted by the abbreviation BR. This approach can be used for estimation of significance at known or assumed s or for estimation of the upper limit of s for predefined α. According to a more recent method, formulated in CERN for standard model Higgs boson search at the LHC [22, 23], s_{\text{ref}} maximizes \( P(\vec{n}|s_{\text{ref}}) \), but is restricted by \( s_{\text{ref}} \leq s \), meaning that it is first found in the \([0, \infty]\) interval and, if it is greater than s, it is equated to s. This can be used for the upper limit. By analogy for the lower limit \( s_{\text{ref}} \) can be chosen such that it maximizes \( P(\vec{n}|s_{\text{ref}}) \) and is constrained by \( s_{\text{ref}} \geq 0 \). This can also be used for estimation of significance, if calculated with \( s = 0 \). In this case the constraint duplicates the constraint assumed at the calculation of the global maximum \( \hat{s} \geq 0 \), so \( s_{\text{ref}} \) is simply equal to \( \hat{s} \). This method will be denoted here as MR, which means the Maximum-Related method.

For instance, for the classical method in the fully binned Poisson case the value of \( Q \) is expressed [7] by

\[
Q = \frac{\prod_{i=1}^{k} e^{-\left(t_{a}a_{i} + t_{b}b_{i}\right)} \left(\frac{a_{i} + b_{i}}{n_{i}}\right)^{n_{i}}}{\prod_{i=1}^{k} e^{-t_{b}b_{i}} (t_{b}b_{i})^{n_{i}}} = e^{-\sum_{i=1}^{k} t_{a}a_{i} s_{i}} \prod_{i=1}^{k} \left(1 + \frac{t_{a}a_{i}s_{i}}{t_{b}b_{i}}\right). \]

(23)

The work [7] also claims that “the confidence in the signal+background hypothesis is given by the probability that the test statistic is less than or equal to the value observed in the experiment, \( Q_{\text{obs}} \):

\[
CL_{s+b} = P_{s+b}(Q \leq Q_{\text{obs}}),
\]

(24)

provided that \( \vec{n} \) that is used at calculation of \( Q \) (at sub-generation, according to our terminology) is distributed according to the signal+background hypothesis, which is indicated by the subscript \( s + b \). “Small values of \( CL_{s+b} \) indicate poor compatibility with the signal+background hypothesis and favor the background hypothesis” [7]. According to the earlier work of Junk [20] \( CL_{s+b} \) is “the confidence level for excluding the possibility of simultaneous presence of new particle production and background (the \( s + b \) hypothesis)”. So this is the usual exclusion of impossible, expressed, for instance, by Eq. (15) for the case of FML and used for setting the upper frequentist limit:

\[
CL_{s+b} = P_{s+b}(Q \leq Q_{\text{obs}}) = \sum_{Q(\vec{n},\vec{a},\vec{b},s)\leq Q_{\text{obs}}} P(\vec{n},s,\vec{a},\vec{b}) = \alpha . \]

(25)

The upper limit is obtained by finding \( s \) that satisfies this equation. The background-related method is not used directly for the lower limit setting, see comments in Ref. [21]. In the maximum-related method the lower limit can be found by the identical formula with the only difference

\[ \text{The report [7] in p. 85 also proposes the use of } s_{\text{ref}} \text{ that maximizes } P(\vec{n}|s_{\text{ref}}) \text{ “in the complete absence of background” and “observation of one or more candidates”} \]

\[ \text{I have not found so far any mentions about the lower limits by this method, so the recipe given above is my extension of this method.} \]

Figure 9: The likelihood ratio method, intervals for 90% one-sided “confidence” (α = 0.1), the uncertainty of the expected background, the known expected signal. The horizontal lines depict optimization by the width of these intervals.
that instead of $0 \leq s_{ref} \leq s$ the condition $s_{ref} \geq s$ should be used at the evaluation of $Q$.

The use of $Q$ for interval setting differs significantly from the ordinary “Neyman construction”, since here the observable test statistic $Q$ depends on the hypothesis about the searched parameter $s$. Therefore instead of vertical lines in the plots $s$ versus $\hat{s}$ like $[C_1, D_1]$ plotted in Fig. one has to consider curved inclined trajectories in the plots $s$ versus $\ln Q(s)$. These trajectories can even cross each other. The picture can be weird enough, but at the absence of nuisance-parameter uncertainties the coverage (possibly conservative) can be proven by similar way as described earlier for FML.

Note that the mentioned above restrictions $s_{ref} \leq s$ and $s_{ref} \geq s$, which are aimed at finding one-sided limits, are the only features that differ the $CL_{s+b}$-based interval finding from the “unified approach” proposed by Feldman and Cousins or for $\vec{n}_\gamma$ distributed according to the background hypothesis:

$$CL_b = P_b(Q \leq Q_{obs}) = \sum_{\vec{n}_\gamma_{s,\bar{a},\bar{b},s}} P(\vec{n}_\gamma | s_u = 0, \bar{a}, \bar{b}).$$

Here $s$ used at the evaluation of $Q$ differs from $s_u$ used for subgeneration, the latter is zero. This $CL_b$ is expected to be close to unity for good signal-like experiments. Note that despite of using the background $\vec{n}_\gamma$, the value of $s$ in Eqs. represents the assumed signal even at the calculation of $CL_b$. It can be the maximum likelihood signal or the signal predicted by theory. The literature describing this background-related method does not offer concrete prescriptions about this. The estimation of significance via $1 - CL_b$ is approximate anyway because this excludes from the $p$-value the probability of obtaining the observed data.

In the maximum-related method corresponding $p$-value is correctly calculated by Eq. with $s = 0$ and with condition $s_{ref} \geq 0$. The value $CL_b$ is used in this method only to correct the upper limit in the case of microscopic signal dependency as described later in this section. In this case $s_u = 0$ and $s$ is a given value, the same as that used at the calculation of $CL_{s+b}$.

At the absence of the nuisance uncertainties the upper limit obtained with $CL_{s+b}$ excludes the true $s$ with probability $\alpha$ even for the experiments microscopically susceptible to the signal. At the application of $CL_{s+b}$ to such experiments there will be strong experiment-by-experiment fluctuations of the reconstructed limit, due to which this limit will be lower than the true $s$ with the probability $\alpha$. This is mathematically correct, but is considered inappropriate in practice. At the unknown nuisance parameters this effect should occur too, although it can have different size. In the opinion of some physicists, if an experiment is not susceptible for the signal, there should not be a way for it to exclude the signal. Moreover, another problem is that the upper limit reconstructed by $CL_{s+b}$ can sometimes be ridiculously low (although never precisely zero, which is mathematically forbidden in this case). This all is in contrast with the upper boundaries of the Bayesian method, of the frequentist treatment of maximum likelihood and of the likelihood ratio method, which all (with slight exception for the latter) provide zero (almost zero for the latter method) probability of non-coverage by the upper limit. These facts are not obvious, but are obtained in calculations.

The limit lower than the true $s$ in such experimental conditions will occur in the cases when the observed statistic is lower than some average or median statistic expected from the background. Provided that the background is calculated correctly, such cases can be interpreted as the downward fluctuations of background. Such experiments are characterized not only by the low $CL_{s+b}$, but also by the low $CL_b$. Dividing $CL_{s+b}$ by $CL_b$ the researcher takes into account how well the experiment is described by the background. The use of $CL_{s+b}/CL_b = \alpha$ instead of $CL_{s+b}$ for search of the upper limit allows one to obtain more conservative limit with 100% coverage for the experiments microscopically weakly susceptible for the signal. See more detailed argumentation in Refs. (7, 20, 21, 22). The confidence-like value $CL_{s+b}/CL_b$ is called $CL_s$ and is traditionally used also as the common name of these methods.

A similar correction for significance has been proposed, but it does not seem to be popular in practice.

We will not correct by $CL_b$ the lower limit by the maximum-related method too.

When the correction by $CL_b$ is used, that is for finding the upper limit, the methods in a whole can be denoted by abbreviations BR–$CL_s$ and MR–$CL_s$. It is not so obvious how to denote the maximum-related method at finding the lower limit and significance. One can either call it MR–$CL_s$, keeping in mind that this is a simple frequentist treatment of the likelihood ratios without the correction by $CL_b$, or call it MR–FLR (Frequency of Likelihood Ratio).

The meaning and interpretation of FLR and/or $CL_s$ methods is very nontrivial. This also affects the programming and calculations. At the implementation of these calculations in software one quickly finds that it is much more convenient to deal with simple $\hat{s}$-values from FMML or FGML than with nontrivial quantities like $Q$, $CL_{s+b}$, $CL_b$ and $CL_s$. MR–$CL_s$ is also very slow in computer calculations.

In the studied test example at the absence of nuisance parameter uncertainties the average upper border position and the coverage appeared to be almost identical with all the other methods.

At the presence of uncertainties the coverage of the intervals obtained by $CL_s$-like methods is not guaranteed “by construction” for any methods of taking this uncertainty into account and it has to be tested numerically.
6.2 The case of unknown $\vec{a}$ and $\vec{b}$

Similarly to FML, there are many combinations. Historically, more combinations have been investigated in this work for the classic BR–$CL_s$ than for MR–$CL_s$. At the parameters given in Section 4.2 $CL_{s+b}$ does not differ much from $CL_s$. As usually, ignoring uncertainties and using $a_i = n_{ai}, b_i = n_{bi}$ does not work well.

Let us first consider the BR–$CL_s$ method.

One can find $\vec{a}$ and $\vec{b}$ that provide the global maximum of $P(\vec{n}, \vec{a}, \vec{b} | \vec{s}, \hat{\vec{a}}, \hat{\vec{b}})$. The subgeneration of $\vec{n}_{s}$ for $CL_{s+b}$ with these $\vec{a}$ and $\vec{b}$ and the subgeneration of $\vec{n}_{a}$ for $CL_b$ with $\hat{\vec{b}}$ do not provide the perfect coverage too, whether one uses the same $\hat{\vec{a}}$ and $\hat{\vec{b}}$ for calculation of $CL_{s+b}$ and $CL_b$ or re-fits them with $\vec{n}_{s}$ of particular subgenerated event and with observed $\vec{n}_{a}$ and $\vec{n}_{b}$.

One can subject the auxiliary measurements $\vec{n}_{a}$ and $\vec{n}_{b}$ to subgeneration too, similarly to SRN–FGML. There are no correctly working methods found that use this approach.

An approach similar to SSP–FGML (see Section 1.2) works well. The subgeneration is carried out by precisely the same way as that for SSP–FGML. From all checked priors only the case with the safe inverse prior for the upper limit was found to work. The lower limit is not calculated in the BR–$CL_s$, see previous section, which is also a problem for the calculation of the significance at unknown nuisance parameters in view of Section 4.3. So only the upper interval border was calculated by this method in this work. This method, similarly to SSP–FGML, has to provide the modeled coverage assuming that the nuisance parameters are distributed by the assumed way. Thus the interpretation exists, but the statistic used is much more complex than the statistic used in SSP–FGML, as was already mentioned in the previous section. The calculations of the studied example show that the upper limit produced by this method for fixed divisions has features similar to that of SSP–FGML.

In the used system of notations this method can be called SSP–GM–BR–$CL_s$. Subgeneration with Safe Priors, Global Maximization, Background-Related $CL_s$.

The same method taken with subgenerated quantities $\vec{n}_{aγ}$ and $\vec{n}_{bγ}$ does not work well.

The marginalization was found to be inappropriate for the case of classic $CL_s$.

One can marginalize the value of $Q$:

$$Q_{\text{full} \text{marg}} = \int \int Q(\vec{a}, \vec{b}) p(\vec{a} | \vec{n}_{a}) p(\vec{b} | \vec{n}_{b}) d\vec{a} d\vec{b} = \int \int \frac{P(\vec{n}, \vec{n}_{a}, \vec{n}_{b} | \vec{s}, \hat{\vec{a}}, \hat{\vec{b}})}{P(\vec{n}, \hat{\vec{b}})} p(\vec{a} | \vec{n}_{a}) p(\vec{b} | \vec{n}_{b}) d\vec{a} d\vec{b}. $$

Here the likelihood $P(\vec{n}, \vec{n}_{a}, \vec{n}_{b} | \vec{s}, \hat{\vec{a}}, \hat{\vec{b}})$ is calculated according to Eq. (3) and analogously $P(\vec{n}, \vec{n}_{b} | \hat{\vec{b}}) = P(\vec{n}, \vec{n}_{b} | \hat{\vec{b}})$, where the elementary probabilities $P(\vec{n}_{a} | \vec{s})$ are calculated by Eq. (6). It seems that this way of full integration is recommended in [20] for marginalization with Gauss functions, with $a_i$ and $b_i$ apparently restricted to be not less than zero. In the case considered here with the Poisson uncertainties for $\vec{a}$ and $\vec{b}$ this is inappropriate because the integral by $db_i$ diverges at $b_i \to 0$ as $b_i^{n_{ai} - n_{bi}}$.

Another possibility of marginalization is the marginalization separately of the numerator and the denominator of likelihood ratio:

$$Q_{\text{marg} \text{by} \text{parts}} = \int \int P(\vec{n}, \vec{n}_{a}, \vec{n}_{b} | \vec{s}, \hat{\vec{a}}, \hat{\vec{b}}) p(\vec{a} | \vec{n}_{a}) p(\vec{b} | \vec{n}_{b}) d\vec{a} d\vec{b} \int P(\vec{n}, \vec{n}_{b} | \hat{\vec{b}}) p(\vec{b} | \vec{n}_{b}) db_i. $$

Numerical investigation shows that at the unknown background no other prior except the inverse one can be used for taking into account the expected background. The subgeneration should also be performed with this prior, by the same method as the subgeneration for FMML, see section 4.2. Similarly to FMML it was assumed that the observed auxiliary measurements are used at the treatment of subgenerated data. However, at the presence of zeros in the expected background histogram the coverage can be lost due to some interruptions in the distributions of $\ln Q(s)$ for the signal in plots $s$ versus $\ln Q(s)$.

Let us now consider MR–$CL_s$. To safe effort, the marginalization technique was not tested here. We also omit some bad tested modifications of the maximization technique.

According to the method described in Refs. [22, 23] the statistic Eq. (22) is replaced by an extended form, which in our notations looks like

$$Q(n, n_{a}, n_{b}, s) = \frac{P(\vec{n}, \vec{n}_{a}, \vec{n}_{b} | \vec{s}, \hat{\vec{a}}, \hat{\vec{b}})}{P(\vec{n}, \vec{n}_{a}, \vec{n}_{b} | \vec{s}, \vec{a}, \vec{b})}, $$

(27)

where, as usually, $\hat{\vec{a}}$ and $\hat{\vec{b}}$ maximize the likelihood for given $s$ and the values $\vec{a}$ and $\vec{b}$ maximize it globally. The values $\hat{\vec{a}}$ and $\hat{\vec{b}}$, obtained by the maximization of $P(\vec{n}, \vec{n}_{a}, \vec{n}_{b} | \vec{s}, \hat{\vec{a}}, \hat{\vec{b}})$ with the observed data for given $s$, including the case $s = 0$ for $CL_b$, are used for the subgeneration in the same way as $\vec{n}_{aγ}$ and $\vec{n}_{bγ}$ in the SARN–FGML method, see Section 4.2 (here we omit the subscript “obs” for briefness). They are used as the parameters of the Poisson distribution Eq. (4) for generation of random $\vec{n}_{aγ}$ and $\vec{n}_{bγ}$. The subgeneration of the main experiment is performed by substituting $\hat{\vec{a}}$ and $\hat{\vec{b}}$ into Eq. (2) and using the result as the parameter of the Poisson distribution for generation of $\vec{n}_{aγ}$. The value $Q(\vec{a}, \vec{n}_{aγ}, \vec{n}_{bγ}, s)$ is used as the threshold. It is compared with $Q(\vec{n}_{aγ}, \vec{n}_{bγ}, \vec{n}_{γ}, s)$ computed by the subgenerated data by Eq. (27) with floating $\hat{\vec{a}}, \hat{\vec{b}}, \hat{\vec{a}}, \hat{\vec{b}}$ and $\vec{a}$ and $\vec{b}$. For example, the right hand side of Eq. (26) is converted into

$$\sum_{\vec{n}_{aγ}, \vec{n}_{bγ}} P(\vec{n}_{γ} | s_a = 0, \vec{a}, \vec{b}) \times P(\vec{n}_{aγ} | \vec{a}) P(\vec{n}_{bγ} | \vec{b}), $$

(28)
where $\hat{a}$ and $\hat{b}$ maximize $P(\hat{a}, \tilde{a}, \tilde{b} | s, \hat{a}, \hat{b})$ at $s = 0$.

Similarly to SARN–FGML, variation of given $s$ leads to simultaneous variation of $\hat{a}, \hat{b}$ and also of $Q(n, n_a, n_b, s)$, $CL_s + b$, $CL_b$, $CL_s$. The total result of this is even more difficult to trace due to the more complex test statistic and the additional correction by $CL_b$. The modeled coverage by construction is not provided. The interpretation of $p$-values assumes single-valued hypotheses for nuisance parameters without taking into account other possibilities. Because of all these issues this method does not seem to be clear and convincing.

This method is denoted here as SARN–MR–$CL_s$. If the correction for $CL_b$ is not applied, it can optionally be denoted by SARN–MR–FLR. The abbreviations should be clear.

For the tested example this method yields good frequentist coverage for the upper limit. The average position of the upper limit is the lowest among all other working (covering) methods, which looks very good, see Fig. [11]. But the coverage of the lower limit is insufficient. This coverage falls below the necessary level at the growth of the numbers of channels. The optimization by forbidding the channels with zero expected background improves the coverage, but only partially. The optimized coverage of both limits separately is about 87% for the test example in the case of the uncertainties of the expected background only. It was explained earlier that such optimization is not applicable as a general method.

The significance obtained by this method can be tested similarly to the tests of SRN–FGML and SARN–FGML methods described in Section [11] and with the use of Eq. [28] with $s = 0$.

The results of such tests indicate that the significance estimates obtained by this method are quantitatively slightly better than the estimates by SSP–FMML and SSP–FGML. However, the average approximate significance can be better than the average “precise” significance, see Fig. [8]. This indicates that the significance is incorrect from the frequentist viewpoint. The maximal difference between the average approximate significance and the average “precise” significance (in the units of $z$ or $\sigma$) for SARN–MR–$CL_s$(FLR) was found to be 0.2 at maximum for the studied example (Fig. [8]) when averaged in $s$ units, or 0.1 in $\sigma$ units at maximum when averaged through $p$-values. Calculation difficulties prevent more comprehensive research, so one cannot guarantee that these differences are always so small or smaller.

The significance estimated by Eq. [28] is true if the values $\hat{a}$ and $\hat{b}$ used in the subgeneration are equal to the real values. The method is very complex and therefore it is not quite clear whether we can yield reasonable result by taking into account a certain fixed values of them and ignoring the other possibilities. This raises the question whether the results supplied by this method depend on the hypotheses about the nuisance parameters, because if the results are independent, it would be unimportant which values of the nuisance parameters are used at subgeneration.

In order to clarify this issue and simultaneously to test another reasonable method an approach similar to SRN–FGML was studied. For the subgeneration stage the nuisance parameters in this method are equated to global maxima, that is to values $\hat{a}, \hat{b}$ maximizing $P(\hat{a}, \tilde{a}, \tilde{b} | s, \hat{a}, \hat{b})$. The rest has to be done precisely as for SARN–MR–$CL_s$. So the difference is only in the nuisance parameters used for subgeneration. This method is called SRN–MR–$CL_s$. The calculations have indicated that the lower confidence limit by this method is even lower and has worse coverage than the lower limit obtained by the SARN–MR–$CL_s$ method. The significance is larger and the difference between the average precise significance and the approximate significance is worse. So the latter method cannot be recommended.

The difference between the results of SRN–MR–$CL_s$ and SARN–MR–$CL_s$ shows that the hypotheses about the nuisance parameters influence the results obtained by the MR–$CL_s$ methods. Furthermore, similarly to the cases of SRN–FGML and SARN–FGML, the dependence on these hypotheses is nontrivial. Without any numerical tests it is hard to tell which method has better frequentist characteristics. Obviously, choosing any specific values of the nuisance parameters for subgeneration one receives strongly
conditional results. In this context the SSP–FMML and SSP–FGML methods look more convincing, since at the subgeneration they take into account the whole assumed distributions representing all possible values of nuisance parameters instead of the arbitrarily chosen fixed values as in the SRN–FGML, SARN–FGML, SRN–MR–CL_s, and SARN–MR–CL methods.

Similarly to FGML, one can foresee that the theoretical uncertainties defined by intervals with uniform distributions inside them can result in the absence of single values of \( \hat{\alpha}, \hat{\beta}, \hat{s}, \hat{a} \) and \( \hat{b} \) in the SRN–MR–CL_s and SARN–MR–CL methods.

There exists an asymptotic approximation of the significance estimate by SARN–MR–CL_s(FLR), see Refs. [22, 23, 37]. In general, the asymptotic values are reached at a large number of observed events. The precision is reportedly good at a wide range of conditions [22], but it is not clear how strongly it can vary at small or zero number of events in separate channels. The asymptotic approximation gives the significance in \( \sigma \)-units as

\[
z = \sqrt{-2 \ln Q(0)}.
\]

This approximation allows one to avoid time-consuming subgeneration-based calculations, which are non-feasible for large significances. The precision of this asymptotic estimate is not clear in particular situations. However, the calculations of the studied test problem indicate that this estimate is very close to that by the explicit calculations by SARN–MR–CL_s(FLR). Therefore one can use this approximation as quickly calculated replacement of the SARN–MR–CL_s(FLR) significance.

Reportedly there exist also asymptotic approximations for SARN–MR–CL_s upper limits [37], but they are not so trivial. In addition, they are less important, since we rarely deal with intervals for \( \alpha \) smaller than 5% and extremely large subgenerated statistics is not required (except for the test of coverage which requires the generation of many “real” experiments and the subgeneration of many for each). Therefore the asymptotic approximations for intervals are not studied in this work.

7 The one-channel “on/off” problem

One particular important comparison with data published in Ref. [12] (an earlier version in Ref. [38]) for the one-channel “on/off” problem with uncertain background is given in table 2. Here \( n_{\text{on}} \) is the same as \( n_1 \), and \( n_{\text{off}} \) is the same as \( n_{\text{off}} \). The significance denoted by \( Z_{\text{BI}} \) is obtained by the test of the ratio of Poisson means, which is reduced to the calculation of \( p \)-value for the binomial distribution. \( Z_{\text{BI}} \) appears to be equal to \( Z_\Gamma \), for which the \( p \)-value is calculated for \( n_1 \) distributed by the pure Poisson distribution with given \( b_1 \). This \( p \)-value is integrated (marginalized) with the uniform prior over the assumed distribution of \( b_1 \). This method is similar but not identical to SSP–FMML and SSP–FGML.

The significance denoted by \( Z_{\text{PL}} \) is the same as the asymptotic approximation to SARN–MR–CL_s(FLR). It appears to be higher than \( Z_{\text{BI}} \), \( Z_\Gamma \) and the significances calculated by SSP–FMML and SSP–FGML. The works [12, 13] studied also the uniformity of the \( p \)-value distribution corresponding to \( Z_{\text{PL}} \). Cousins, Linnemann and Tucker [12] conclude that it has “at most modest under-coverage”, that is the integrated density of the \( p \)-value distribution at the left end is greater than uniform. As explained in the introduction this may be an evidence of the bad method. Section 6.2 proves this in a different way. Therefore relatively high estimates of significance by this method are not reliable.

As seen from the table, the methods SEP–FMML, SEP–FGML, SRN–FGML, SARN–FGML and SRN–MR–CL_s provide higher significances as well. But previous sections prove that they are not reliable either.

8 Conclusions

Good features are exhibited by three related methods: by the method of modified Bayesian central intervals and by two frequentist methods based on the frequentist treatment of the maximum likelihood estimate. The latter two methods are denoted here by the abbreviations SSP–FMML and SSP–FGML. All three methods use the so-called safe priors for reconstruction of the distributions of the expected signal and expected background.
The optimization of divisions is necessary for these three methods in order to avoid divisions, by different reasons, with too few and too many channels. The general way of optimization for these three methods is the selection of the division that produces the shortest interval from those for which there are no zeros in the distributions of the expected signal. The optimization for the SSP–FMML and SSP–FGML methods is better to perform with the modified central Bayesian intervals.

The intervals optimized by this way are reasonably short and have reasonable probability of covering the true value. Thus, their frequentist coverage is approximately equal or greater than stated. In addition, the boundaries of the SSP–FMML and SSP–FGML intervals provide the coverage by construction for a certain non-frequentist sequence of experiments in which the parameter of interest is fixed at the true value as assumed in the frequentist approach, and the nuisance parameters fluctuate as prescribed by the Bayesian analysis of the auxiliary measurements. Whether the frequentist coverage is provided or not, this “modeled coverage”, which is always present in these methods, is an important feature itself. It can also be understood as the interpretation of the confidence limits, clear and convincing in the case of SSP–FMML and SSP–FGML. The same interpretation holds for the p-values obtained by these methods. The modified Bayesian intervals have the Bayesian interpretation in addition to the approximate frequentist coverage.

The intervals reconstructed by all three methods are similar in the average. They are not the shortest in the average among the intervals obtained by some other FML-based and likelihood ratio-based methods. However, all the alternative methods that seem to produce slightly shorter intervals have problems with the frequentist coverage.

Among the frequentist methods the SSP–FMML method can be considered as the most general method, because it is assumed to be useful in the widest range of problems. If the SSP–FGML method is applicable, it can have slightly better performance than the SSP–FMML method.

In addition to good confidence intervals the SSP–FMML and SSP–FGML methods yield very reliable significance estimates, though these estimates are not the most optimistic, as well as the intervals obtained by these methods.

It is not suggested that one of these or other methods is the best. Because of diverse features exhibited by different methods, the choice is thought to depend also on priorities and circumstances, such as the preference of the Bayesian or frequentist analysis in a given field of research, feasibility of calculations, etc.

The likelihood ratio (profile likelihood) method provides optimistic intervals which usually do not have the precise frequentist coverage. This conclusion is in agreement with other reports. However, this method is known to be implemented very easily. Therefore it can be used for fast approximate estimates of the intervals.

The frequentist treatment of the likelihood ratios is implemented in the methods known as the $CL_s$ methods, from which the most advanced is the LHC-style $CL_s$ method, denoted here as SARN–MR–CL$_s$ or SARN–MR–FLR in the case of no correction by $CL_b$. It provides very small intervals, but their frequentist coverage can be slightly smaller than stated, both for fixed divisions and for optimization by the interval width. There is no modeled coverage, or, in other words, there is no model for nuisance parameters at which the coverage is provided by construction. The interpretation of $p$-values assumes only single-valued hypotheses for nuisance parameters without taking into account other possibilities. There is no satisfactory method of optimization found for this method. This method provides very small $p$-values and hence large significance estimates, but, similarly, their frequentist features are imperfect. In short, the significance obtained by this method in a sequence of experiments can appear to be slightly greater in average than the true average significance. The full method is very complex for implementation. However, its asymptotic approximation is known to be very simple. It is confirmed that the asymptotic estimates of the significance are close to the estimates obtained by the full calculations according to this method.

References

[1] R. D. Cousins, Am. J. Phys. 63 (1995) 398.
[2] G. J. Feldman, R. D. Cousins, Phys. Rev. D 57 (1998) 3873.
[3] A. Stuart, J. Keith Ord, S. Arnold, Kendall’s advanced theory of statistics, vol 2A, Classical Inference and the Linear Model. 6-th ed. 1999, Oxford University Press Inc.
[4] G. D’Agostini, Bayesian reasoning versus conventional statistics in high energy physics arXiv:physics/9811046v1, 1998.
[5] M. Mandelkern J. Schultz, J. Math. Phys. 41 (2000) 5701.
[6] R. D. Cousins, Treatment of nuisance parameters in high energy physics, and possible justifications and improvements in the statistics literature, Proceedings of PHYSTAT 05 http://www.physics.ox.ac.uk/phystat05/proceedings/
[7] A. L. Read, Modified frequentist analysis of search results (the CL$_s$ method), 1st Workshop on Confidence Limits, CERN, Geneva Switzerland, 2000, CERN-2000-005, p. 81.
[8] J. Heinrich, Review of the Banff Challenge on Upper Limits, PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics, CERN, Geneva, Switzerland, 2007, http://cdsweb.cern.ch/record/1021125
[9] T. Juck, Banff Challenge 2, Proceedings of the PHYSTAT 2011, http://cds.cern.ch/record/1306523/files/CERN-2011-006.pdf
[10] W. A. Rolke, A. M. Lopez, J. Conrad, Nucl. Instr. and Meth. A551 (2005) 493.
[11] F. Tegenfeldt, J. Conrad On Bayesian Treatment of Systematic Uncertainties in Confidence Interval Calculation, Nucl. Instr. and Meth. A539 (2005) 407.
[12] R. D. Cousins, J. T. Linnemann, J. Tucker, Nucl. Instr. and Meth. A595 (2008) 480.
[13] K. Cranmer, Statistical challenges for searches for new physics at the LHC, Proceedings of PHYSTAT 05, arXiv:physics/0511028v2, 2006.
[14] K.S. Cranmer Frequentist Hypothesis Testing with Background Uncertainty, PHYSTAT 2003, SLAC, September 8-11, 2003, arXiv:physics/0310108v1, 2003.
[15] R. D. Cousins, Advanced Statistics for High Energy Physics, Fourth CERN-Fermilab Hadron Collider Physics Summer School, June 8, 2009, http://indico.cern.ch/getFile.py/access?contribId=1&resId=0&materialId=slides&confId=44587
[16] A. O’Hagan and J. Forster, Kendall’s advanced theory of statistics, vol 2B, Bayesian Inference, 2-th ed. 2004, John Wiley & Sons, Ltd.
[17] J. Heinrich et al., Interval estimation in the presence of nuisance parameters. 1. Bayesian approach, September 27, 2004, CDF/MEMO/STATISTICS/PUBLIC/7117, arXiv:physics/0409129
[18] G. Choudalakis, How to Use Experimental Data to Compute the Probability of Your Theory, arXiv:1110.5295v1, 2011.
[19] S. Ciampolillo, Nuono Cimento A111 (1998) 1415.
[20] T. Junk, Nucl. Instr. and Meth. A434 (1999) 435.
[21] A. L. Read, J. of Phys. G 28 (2002) 2693.
[22] The ATLAS Collaboration, The CMS Collaboration, The LHC Higgs Combination Group, Procedure for the LHC Higgs boson search combination in Summer 2011. August 2011, ATL-PHYS-PUB-2011-11, CMS NOTE-2011/005.
[23] CMS Collaboration, Phys. Let. B710 (2012) 26.
[24] A. Stuart and J. K. Ord, Kendall’s advanced theory of statistics, vol 1, Distribution Theory, 6-th ed. 2000, Oxford University Press Inc.
[25] J. Neyman, Phil. Trans. Roy. Soc. of London, A236 (1937) 333.
[26] G. D’Agostini, About the proof of the so called exact classical confidence intervals. Where is the trick? arXiv:physics/0605140v2, 2010.
[27] R. E. Kass, L. Wasserman, J. Am. Statist. Assoc. 91 (1996) 1343.
[28] Discussion with Jose Bernardo on Bayesian reference analysis, Transcribed and edited by Luc Demortier. Proceedings of the PHYSTAT 2011, http://cds.cern.ch/record/1306523/files/CERN-2011-006.pdf
[29] J. Heinrich, The Bayesian approach to setting limits: what to avoid, Proceedings of PHYSTAT 2005, http://www.physics.ox.ac.uk/phystatt/proceedings/
[30] The program TEfficiency::ClopperPearson() from the ROOT system: R. Brun and F. Rademakers, Nucl. Instr. & Meth. A 389 (1997) 81.
[31] T. Müller, J. Ott, J. Wagner-Kuhr, Theta — a framework for template-based modeling and inference, Karlsruhe Institute of Technology, June 17, 2010, http://www-ekp.physik.uni-karlsruhe.de/~ott/theta.html/index.html
[32] R. D. Cousins, V. L. Highland, Nucl. Instr. and Meth. A320 (1992) 331.
[33] G. E. P. Box, J. Roy. Statist. Soc. A143 (1980) 383.
[34] J. De La Horra, M. T. Rodríguez-Bernal, Rev. R. Acad. Cienc. Exact. Fis. Nat. (Esp) 93 (1999) 321.
[35] I. B. Smirnov, LikelihoodLib — Fitting, Function Maximization, and Numerical Analysis, Preprint PNPI–2421, 2001.
[36] J. Neyman and E. S. Pearson, Phil. Trans. A 231 (1933) 289.
[37] G. Cowan, K. Cranmer, E. Gross, O. Vitells, Eur. Phys. J. C71 (2011) 1554.
[38] J. T. Linnemann, Measures of Significance in HEP and Astrophysics, Proceedings of PHYSTAT 2003.