PUBLISHED BY INSTITUTE OF PHYSICS PUBLISHING FOR SISSA

RECEIVED: *April 1, 2006*
REVISED: *December 23, 2006*
ACCEPTED: *December 28, 2006*
PUBLISHED: *January 22, 2007*

Computation of confidence levels for search experiments with fractional event counting and the treatment of systematic errors

Peter Bock

Physikalisches Institut der Universität Heidelberg,
Philosophenweg 12, D69120 Heidelberg, Germany
E-mail: peter.bock@physi.uni-heidelberg.de

ABSTRACT: A method is described which computes, from an observed sample of events, upper limits for the production rate of new particles, or, for the case of an observed excess of events over background, the probability for an upward fluctuation of the background. It is based on weighted event counting depending on a discriminating variable. Candidates may be produced in different reaction channels with different detection efficiencies and different background. Systematic errors with arbitrary correlations are taken into account in the confidence level calculations. In addition, they are are incorporated in the weight definition. Conditions under which the Bayesian and the frequentist treatment of systematic errors give identical results are derived. It is shown that the significance of an observation of a signal is generally overestimated in low statistics experiments. Simple approximate formulae for observed and expected confidence levels are given for the limiting case of high count rates. A special statistical test of a given signal-to-background-ratio using the distributions of the discriminating variable and fixing the total theoretical intensity to the observed number of events, is described.

KEYWORDS: Statistical Methods, Higgs Physics.

Contents

1. Introduction **2**

2. Specification of the weight function **2**

2.1 Discriminating variables 2

2.2 The computation of event weights 3

2.3 Two discriminating variables 8

2.4 Related approaches 8

3. Weight distributions **9**

3.1 Folding procedure 9

3.2 Some analytic results 12

4. Applications **13**

4.1 Upper limits without background subtraction 13

4.2 Upper limits with background subtraction 15

5. Systematic errors **18**

5.1 Parameterization of systematic errors 18

5.2 Correction of confidence levels in the frequentist approach 19

5.3 Equivalence between the frequentist and the Bayesian treatment of systematic errors 20

5.4 Numerical treatment of systematic errors 24

5.5 Poisson distribution at small rates 24

6. Event weighting with systematic errors **26**

7. Confidence levels from the shapes of distributions **29**

8. Summary **32**

A. Optimal weights depending on systematic errors **33**

B. Comments on the comparison of two arbitrary hypotheses **34**

1. Introduction

The analysis of particle search experiments and the interpretation of the results can be quite complex. In many cases several physical channels with different systematic errors and even different experiments have to be combined. Sophisticated and efficient event taggers have been developed to detect specific event topologies. If an excess over the expected background is observed in the data, an immediate question is to what extent an upward fluctuation of the background can be ruled out. These issues have been discussed by a variety of authors (see the summaries given in refs. [1] to [5], and refs. [6] to [8]).

In this paper the method of fractional event counting is described which uses, as the indicator for a signal, a weighted sum over the observed events. The weights, also called the filter function, are computed from physical variables of the candidates. The method was originally applied in Higgs search experiments (refs. [9] to [11]) but was only briefly described [11]. It is the scope of this paper to give a more detailed description, as well as a presentation of recent developments. The notations and definitions in this paper follow widely the conventions of the LEP Higgs working group [15]. They are recapitulated in section 2.1 and the first part of section 2.2.

Originally, the weight definition was introduced as a heuristic approach. In section 2.2 a more general formula for the weight computation is derived from first principles, using various criteria to optimize the signal sensitivity. Similarities and differences to the likelihood ratio method are pointed out. After a recapitulation of numerical algorithms in section 3, some numerical examples for the calculations of of upper limits are given in section 4. A comparison with other methods for the case of unknown background is also presented there.

Sections 5 and 6 of this paper discuss the treatment of systematic errors. For complex analyses these are commonly included following the method of ref. [12]. This is a Bayesian method, superimposed on a frequentist approach to compute confidence levels. It is investigated under which conditions this method agrees with a frequentist ansatz. A fast method to include symmetric systematic errors with arbitrary correlations into confidence level computations is presented. In addition, a very simple example of an incorrect treatment of systematic errors is given. In section 6 it is shown how the sensitivity for signal detection can be improved by including systematic errors in the weight definition.

The detection of a signal is based partly on event counting and partly on different shapes of the signal and background weight distributions. To distinguish between these effects, a special version of fractional event counting is suggested, in which the theoretical number of events is normalized to the observed number.

2. Specification of the weight function

2.1 Discriminating variables

The aim of any statistical analysis in a search experiment is to distinguish between two physical hypotheses:

(A) The data consist of background only,

(B) the data consist of background plus a hypothetical signal.

In such analyses the observed events are ordered according to their signal likeness, given by the value of a discriminating variable ξ, which is computed for each event. This variable can be a reconstructed mass, a likelihood computed from several physical observables or the result of a neural network analysis. It is assumed that theoretical predictions for the spectral shapes of signal and background, $s(\xi) \geq 0$ and $b(\xi) \geq 0$, exist. Data may be available for several decay modes of a new hypothetical particle and several accelerator energies, and may come from more than one measurement. These different cases are referred to as experimental channels below. The variable ξ may vary from channel to channel. Instead of ξ, an arbitrary monotone function of it could be used as well. It is well known, and will be an automatic result of the next subsection, that the final results are independent of such a redefinition, apart from binning effects.

All histogram bins are assumed to be statistically independent. It is therefore not allowed that an event enters the analysis twice. If a channel overlap exists, for instance between two final states, the two corresponding channels must be rearranged into three: the exclusive selection of events in the two original channels and the overlap between the two with a new definition of ξ.

In most cases the spectra of ξ are available in form of Monte Carlo histograms $b_{ki} = b(\xi_{ki})$ for the background and $s_{ki} = s(\xi_{ki})$ for the signal. Here, the index k is used to identify a channel and i indicates the value of its discriminating variable. The trivial case of simple event counting corresponds to the limitation to one histogram bin. Throughout this paper it is assumed that the sum over all histogram contents is normalized to the expected event rate. For later use, signal efficiencies per bin are defined as

$$\epsilon_{ki} = \frac{s_{ki}}{r},$$

where $r = \sum_{ki} s_{ki}$ is the total signal rate. Branching ratios of decays, channel dependent cross sections and different luminosities are incorporated into ϵ_{ki}.

2.2 The computation of event weights

Using the s_{ki} and b_{ki} histograms, event weights w_{ki} can be defined. Their definition is, however, not unique. Different choices do not give the same result for a given experiment. The weights can be chosen, however, so as to optimize the expected discrimination between hypotheses (A) and (B).

If the w_{ki} are known, the total weight of an event sample, commonly called 'test statistic' X, is defined as

$$X = \sum_{\text{events } l} w_{k(l)i(l)} \ .$$

The sum extends over all candidates of an experimental data set or a Gedanken experiment. The indices $k(l)$ indicate the channels and $i(l)$ are the ξ bins to which the events belong.

If an experiment is repeated many times, the resulting total weights show statistical fluctuations. They are described by probability density functions $P_b(X)$ and $P_{sb}(X)$. These functions refer to the hypotheses (A) (background only) and (B) (background plus signal). They are related to the input histograms s_{ki} and b_{ki} and depend on the weight definition. Implicitly they depend on the total signal and background rates. Their computation will be described in detail in the next section.

The hypothesis testing is based on the two confidence levels

$$CL_b(X_{\rm cut}) = \int_0^{X_{\rm cut}} P_b(X)dX \quad \text{and} \quad CL_{sb}(X_{\rm cut}) = \int_0^{X_{\rm cut}} P_{sb}(X)dX \ . \tag{2.1}$$

They are the probabilities that the test statistic X is smaller than or equal to $X_{\rm cut}$ [13]. When the cut $X_{\rm cut}$ is equal to the test static X_{obs} observed in an experiment, a small value of $CL_{sb}(X_{obs})$ indicates a measured deficit with respect to hypothesis (B) and it is said that hypothesis (B) is ruled out with the confidence $1 - CL_{sb}(X_{obs})$. Similarly, a value of $CL_b(X_{obs})$ close to unity indicates an excess over the background with an expected probability $1 - CL_b(X_{obs})$ in the absence of a signal. The previous definitions and the nomenclature were adopted by the LEP Higgs working group [10, 11].

According to the central limit theorem, the functions P_{sb} and P_b, in the limit of high rates, can be approximated by Gaussians.

$$\begin{aligned} P_b(X) &= \frac{1}{\sqrt{2\pi}\sigma_b} \exp\left(- \frac{(X- < X >_b)^2}{2\sigma_b^2} \right); \\ P_{sb}(X) &= \frac{1}{\sqrt{2\pi}\sigma_{sb}} \exp\left(- \frac{(X- < X >_{sb})^2}{2\sigma_{sb}^2} \right). \end{aligned} \tag{2.2}$$

The expectation values of X are given by

$$< X >_b = \sum_{k,i} w_{ki} b_{ki} \ ; \quad < X >_{sb} = < X >_s + < X >_b = \sum_{k,i} w_{ki}(r\epsilon_{ki} + b_{ki}) \ . \tag{2.3}$$

The sums extend over all channels k and ξ bins i. The variances of the X distribution due to statistical fluctuations of the event numbers and the ξ values are given by

$$\sigma_b^2 = \sum_{k,i} w_{ki}^2 b_{ki} \ ; \quad \sigma_{sb}^2 = \sigma_s^2 + \sigma_b^2 = \sum_{k,i} w_{ki}^2 (r\epsilon_{ki} + b_{ki}) \ . \tag{2.4}$$

There is no unique criterion to discriminate between the hypotheses (A) and (B). In this paper the following optimization strategies are considered:

(i) The mean confidence level $< CL_{sb} >_b$ for the interpretation of an arbitrary test statistic X from the background source (A) as signal plus background (B) should be minimized.

(ii) The mean confidence level $< CL_b >_{sb}$ for interpretation of an arbitrary test statistic X from the combined signal and background source (B) as background (A) should be maximized.

(iii) For the case in which a signal is observed, the probability for an upward fluctuation of the background test statistic to the median signal plus background level should be minimized.

(iv) Criterion (iii) is generalized by assuming that the background fluctuates to a signal plus background level different from its median value by K standard deviations.

(v) The probability to find an existing signal should be maximized.

(vi) The measurement of a hypothetical signal rate should have the lowest statistical error.

(vii) For the case of no signal, the computed bounds n_{CL} on the total signal rate r, at a given confidence level CL, should be minimized.

Not all of the above requirements are equivalent and the resulting weights differ. These will be computed in the high rate limit.

Criteria (i) and (ii): Overlap of the test statistic for data samples from the sources (A) and (B). It is well known that requests (i) and (ii) are identical: Criterion (i) uses simply the mean probability that an arbitrary Gedanken experiment with signal and background events has a total weight smaller than or equal to the weight of an arbitrary experiment counting background only. The equivalence of (i) and (ii) follows from the fact that the probability in (ii) is complementary.

The probability densities at $X = 0$ are negligible and one finds, with equations (2.1) to (2.4),

$$\begin{aligned} < CL_{sb} >_b = & \frac{1}{\sqrt{2\pi}\sigma_b}\int_{-\infty}^{\infty} dX \cdot \exp\left(- \frac{(X - < X >_b)^2}{2\sigma_b^2}\right) \cdot \\ & \frac{1}{\sqrt{2\pi}\sigma_{sb}}\int_{-\infty}^{X} dY \cdot \exp\left(- \frac{(Y - < X >_{sb})^2}{2\sigma_{sb}^2}\right). \end{aligned}$$

The brackets on the left hand side indicate the statistical mean value. Both physical models appear in the equation. The events consist of background, which is indicated by the index 'b' outside the brackets but they are analyzed in terms of signal and background (CL_{sb}). The double integral can be simplified to

$$< CL_{sb} >_b = \frac{1}{\sqrt{2\pi \cdot (\sigma_{sb}^2 + \sigma_b^2)}} \int_{-\infty}^{-<X>_s} dZ \cdot \exp\left(- \frac{Z^2}{2(\sigma_{sb}^2 + \sigma_b^2)}\right) , \tag{2.5}$$

where

$$< X >_s = \sum_{k,i} w_{ki} s_{ki} \; .$$

is the expectation value of X for signal events. The probability $< CL_{sb} >_b$ depends on the ratio $< X >_s / \sqrt{\sigma_{sb}^2 + \sigma_b^2}$ only, to be maximized. Since a common scale factor in all w_{ki} cancels out in the confidence levels, the mean value $< X >_s$ can be fixed. The optimization criterion is then, with a Lagrangian factor λ,

$$\frac{\partial(\sigma_{sb}^2 + \sigma_b^2)}{\partial w_{ki}} - \lambda \frac{\partial < X >_s}{\partial w_{ki}} = 0 \; .$$

This equation can be solved for w_{ki}. The solution is proportional to λ. After a renormalization the result can be written as

$$w_{ki} = \frac{s_{ki}}{s_{ki} + 2b_{ki}} \ . \tag{2.6}$$

The factor 2 appears because the width of the background distribution enters twice.

Criterion (iii): Minimal upward fluctuation of the background, median level. The ratio $< X >_s / \sigma_b$ between the mean signal test statistic and the background error has to be maximized.

Criterion (iv): Minimal upward fluctuation of the background, general case. In the high rate limit, the expected probability for an upward fluctuation of the background is given by

$$1 - E[CL_b]_{sb} = 1 - \frac{1}{\sqrt{2\pi}\sigma_b} \int_{-\infty}^{X_{\text{cut}}} dX \cdot \exp\left(- \frac{(X - < X >_b)^2}{2\sigma_b^2} \right) \ . \tag{2.7}$$

The function 'E' is an expected value for the variable within the brackets, and the indices outside the brackets indicate the source of the events. The integration limit X_{cut} is given by the requirement that the test statistic for signal plus background deviates from its median value by K standard deviations

$$X_{\text{cut}} = r \sum_{k,i} \epsilon_{ki} w_{ki} + < X_b > - K \cdot \sigma_{sb} \ .$$

In order to determine the weights, the ratio

$$\frac{(r \sum_{k,i} \epsilon_{ki} w_{ki} - K\sigma_{sb})^2}{\sigma_b^2}$$

has to be maximized.

Criterion (v): Maximal probability to detect a signal. The maximum chance to detect a signal is obtained by minimizing the expected probability for a downward fluctuation below a test statistic X_{cut}, computed for background

$$E[CL_{sb}]_b = \frac{1}{\sqrt{2\pi}\sigma_{sb}} \int_{-\infty}^{X_{\text{cut}}} \exp\left(- \frac{(X - < X >_{sb})^2}{2\sigma_{sb}^2} \right) \ . \tag{2.8}$$

If the background level differs from the median value by K standard deviations, one has

$$X_{\text{cut}} = < X >_b + K \cdot \sigma_b \ .$$

The weights w_{ki} are given by maximizing the ratio

$$\frac{(\sum_{k,i} r\epsilon_{ki} w_{ki} - K\sigma_b)^2}{\sigma_{sb}^2} .$$

Criterion (vi): Best measurement of the signal rate. The measurement of a hypothetical signal rate is most significant, if the ratio

$$< X >_s^2 / \sigma_{sb}^2$$

is maximal. This request is contained as the special case $K = 0$ in criterion (v).

The functional form of w for the cases (iii) to (vi) is obtained in the same way as equation (2.6). Fixing the sum $\sum_{k,i} w_{ki}\epsilon_{ki}$ to set the w_k scale, computing the derivatives with respect to w_{ki} and absorbing all $k-$ and i-independent sums into common constants, results for all cases, after a final renormalization, in the functional form

$$w_{ki} = \frac{\mathcal{N} R \epsilon_{ki}}{R\epsilon_{ki} + b_{ki}} \ . \tag{2.9}$$

The confidence levels are invariant against multiplication of all w_{ki} by a common factor. A normalization constant $\mathcal{N}$ is introduced to adjust the overall maximum weight to 1, but this factor could also be dropped. The general result (2.9) thus depends on one rate parameter R only, which has to be tuned to fulfill one of the optimization criteria. To guarantee a positive denominator in all cases, R should be positive. In general, R is not equal to the signal rate r but is proportional to it. Equation (2.6), for criteria (i) and (ii), is contained as the special case $R = r/2$. For criterion (iii), i.e. the observation of a signal at its median value and a minimum upward fluctuation of the background, the result R approaches zero. This means that the weight is proportional to the signal-to-background-ratio. For condition (vi) one obtains $R = r$. In the cases (iv) and (v) R depends on K.

Criterion (vii): Minimal rate limit for the signal. To find the lowest expected upper limit n_{CL} for a non-existing signal, a certain number of standard deviations K equivalent to CL has to be introduced. A fixed value K is equivalent to a cut in the signal plus background distribution of X at $X_{\text{cut}} = r \cdot \sum_{k,i} \epsilon_{ki} w_{ki} + < X >_b - K \cdot \sigma_{sb}$. It is assumed that the background is observed at its median level. The rate limit fulfills then the equation

$$\sum_{k,i} n_{CL}\epsilon_{ki} w_{ki} - K\sigma_{sb} = 0 \ .$$

The error σ_{sb} depends on n_{CL} implicitly. Differentiating the last equation with respect to w_{ki} and setting $dn_{CL}/dw_{ki} = 0$ gives equation (2.9) again with $R = n_{CL}$. This is a self consistency relation between the expected rate limit and the parameter R.

The weights (2.9) depend on the ϵ_{ki} to b_{ki} ratio only and are therefore invariant against ξ transformations, which rescale both distributions with the same ξ dependent factor. Equation (2.9) was derived in the high rate limit. If applied to low rates, it is not anymore optimal but is still very close to the optimum and gives still bias free results. Of course, the simple analytic formulae (2.5), (2.7) and (2.8) for the confidence integrals and the results for the R values given here are then not anymore valid.

Throughout this paper it is understood that the weight algorithm, including the parameter R, is fixed *a priori* and not fitted to observed data. This makes it necessary to generalize the criteria (i) to (vii) to non-Gaussian distributions and to compute the functions P_{sb}, P_b and the expected confidence levels $< CL_{sb} >_b, E[CL_{sb}]_b$ and $E[CL_b]_{sb}$ numerically, using theoretical predictions for ϵ_{ki}, b_{ki} and r. The parameter R has to be varied until a chosen optimization criterion is fulfilled.

As will be shown later, the optimization procedure allows variations of R within rather wide regions if the user allows relative numerical tolerances of the order of a per mille for the expected confidence levels or expected rate limits. On the contrary, for a specific data set the results may be R dependent. In general, this effect is small at large rates. However, in low statistics experiments the analyses may become rather ambiguous.

A user has to select a parameter R without introducing subjectivity. In many cases the signal-to-background-ratio is a suitable choice. This is especially true if a signal is observed but no theoretical prediction for the cross section exists. An expected signal rate is not needed to define w and the function P_b can be used to compute the probability for an upward fluctuation of the background to the measured test statistic. If a definite signal prediction has to be checked, the value $R = r/2$ is the appropriate choice. For the determination of upper bounds the expected limit $E[n_{CL}]$ can be minimized. An example for this procedure is given in section 4.2. This strategy works if the background is sufficiently large.

2.3 Two discriminating variables

Experiments searching for a new hypothetical particle often use a signal likelihood variable to reduce the background, and, in many cases, the likelihood definition does not contain the reconstructed particle mass explicitly. The distribution of the reconstructed masses m is only weakly correlated to the distribution of the likelihood L, and an overall discriminating variable ξ can be constructed. Following equation (2.6), a simple product ansatz can then be used:

$$\xi = \frac{D_{sm}(m) D_{sL}(L)}{D_{sm}(m) D_{sL}(L) + 2 D_{bm}(m) D_{bL}(L)}\,, \tag{2.10}$$

where the D's are the probability density functions and the indices indicate the physical observables and signal (s) or background (b). This procedure was used in Higgs searches of the OPAL collaboration [14].

The definition (2.10) has the property that the weights w_{ki}, computed from a Monte Carlo sample with equation (2.6), agree with ξ if the two physical variables are truly uncorrelated, i.e. the product ansatz is correct. Any deviation indicates the presence of correlations or unacceptably large fluctuations in the Monte Carlo samples used to generate the histograms. This was found in the analysis of ref. [14], where the initial observation of a few statistical anomalies made additional Monte Carlo simulations necessary.

2.4 Related approaches

An alternative approach, which is used quite often, is the ordering of experiments according to the likelihood ratio L_{sb}/L_b between the signal plus background and the background interpretation of a data set [13, 6, 7]). Poisson statistics give

$$L_{sb}/L_b = \exp(-r) \frac{\prod_{k,i} (s_{ki} + b_{ki})^{n(k,i)}}{\prod_{k,i} b_{ki}^{n(k,i)}}\,, \tag{2.11}$$

where $n(k,i)$ is the number of candidates observed in the bin combination (k,i). It is known that this likelihood-ratio-method is equivalent to event counting with a weight

$$w_{ki} = \ln\left(1 + \frac{s_{ki}}{b_{ki}}\right) . \tag{2.12}$$

The power expansion in terms of the signal-to-background ratio is

$$w_{ki} = \frac{s_{ki}}{b_{ki}} - \frac{1}{2}\frac{s_{ki}^2}{b_{ki}^2} + \frac{1}{3}\frac{s_{ki}^3}{b_{ki}^3} + \dots \quad .$$

This can be compared with twice the expansion of equation (2.6). It turns out that the first two terms agree and the difference of the third terms is $s_{ki}^3/(12 \cdot b_{ki}^3)$ only so that the results of both methods are very similar if $s_{ki}/b_{ki} \lesssim 1$.

Significant differences between the two approaches are possible if one or more candidates are present in phase space regions where $s_{ki} \gg b_{ki}$. If the background level is correct, the appearance of only one candidate can be a significant indication for a signal. However, in many cases the background is underestimated and the singularity in equation (2.12) then can produce a spurious discovery. Such an effect introduced by one candidate, probably caused by underestimated background, was found in an earlier LEP combination of Higgs searches [11]. It had no impact on the final result, however, because the candidate mass was well above the combined mass limit, where the theoretical signal cross section was too small for a signal interpretation.

Contrary to equation (2.12), equation (2.9) approaches a constant event weight in the limit $b_{ki} \to 0$ and is thus robust against such effects. Another important advantage of (2.9) is that it can be generalized to incorporate systematic errors, which are correlated between the ξ bins (see section 6).

Definition (2.9) is related to the maximum likelihood fit of the signal rate. The logarithmic derivate of the likelihood is

$$\frac{d\ln L_{sb}}{dr} = \frac{1}{L_{sb}} \cdot \frac{dL_{sb}}{dr} = \frac{X}{r} - 1\,,$$

with

$$X = \sum_l \frac{\epsilon_{k(l)} \cdot r}{\epsilon_{k(l)} \cdot r + b_{k(l)}} \ .$$

which is equivalent to (2.9) with $R = r$. The likelihood fit determines r from the condition $X = r$.

3. Weight distributions

3.1 Folding procedure

After the weight function $w(\xi)$ has been specified, the density distributions $D(\xi)$ can be transformed into distributions of w, for one event denoted by $P_1(w)$. The symbol D stands for s or b. The histogram conversion is illustrated in figure 1. The cumulated integral

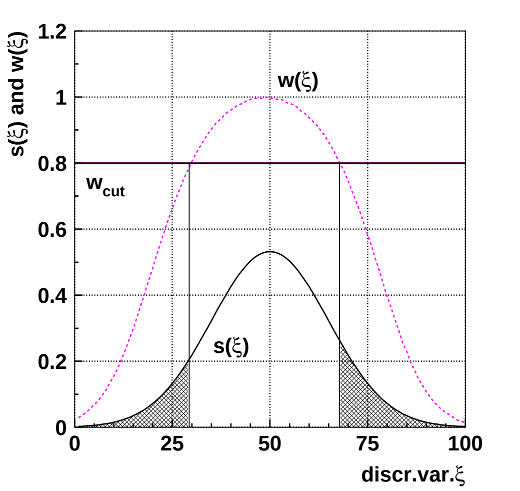

Figure 1: Construction of the cumulated weight distribution of signal events from their ξ distribution and the weight function $w(\xi)$.

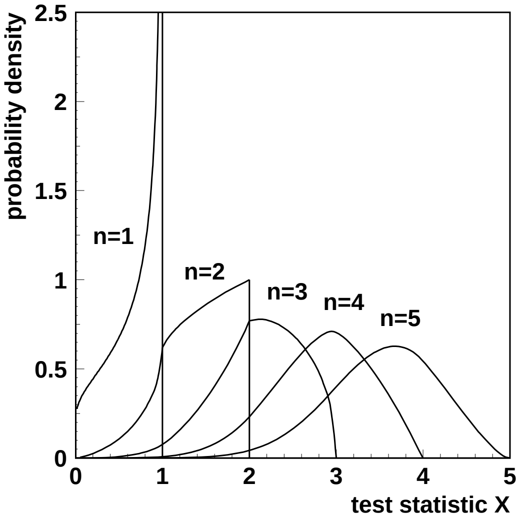

Figure 2: Spectra of the test statistic X for fixed numbers of events. The distributions are for a small signal-to-background ratio and a Gaussian signal over a constant background. The functions are given for the signal.

$\int_0^{w_{\rm cut}} P(w)dw$, for a certain value $w_{\rm cut}$, is illustrated by the shadowed area. In case of histograms, all ξ bins with $w_{ki} \leq w_{\rm cut}$ have to be counted. The cumulated spectrum can be converted into the differential one by taking bin-to-bin differences. The central w values of the bins are assigned to all predicted and observed events in that bin. The analytic formula for a continuous function is

$$P_1(w) = \sum_l \frac{D(\xi_l)}{|\frac{dw}{d\xi}(\xi = \xi_l)|} \; . \tag{3.1}$$

The sum extends over all solutions ξ_l of the equation $w(\xi) = \xi$ and appears because the backward transformation from w to ξ is not unique.

The differential histograms $P_1(w_j)$ may have gaps, but these are never populated by Monte Carlo or data events. The extreme case would be a delta function at $w = 1$ for simple event counting. Since the distributions are not constant within a bin, binning effects can introduce relative errors in the rate limits of the order of $1/(< w > \cdot$ number of w bins).

The distribution of the test statistic $X = \sum_{l=1}^{n} w_{k(l)i(l)}$, for a fixed number of n events, can now be computed from the distribution for one event by iterative folding:

$$P_n(X) = \int_{\max(0,X-(n-1))}^{\min(1,X)} P_{n-1}(X - w) \cdot P_1(w) \cdot dw \; . \tag{3.2}$$

The maximum weight for one event was normalized to 1. The integration limits guarantee that the arguments do not become negative or exceed their upper limits. In general, these equations have no analytic solutions and must be evaluated numerically by matrix multiplication. The stepwise evolution of P_n, for a Gaussian signal and constant background, is shown in figure 2. The singularity of P_1 at $X = 1$, due to the maximum in the ξ distribution, survives as a step for $n = 2$ and as a vertical slope at the upper end for $n = 3$. At $n = 4$ all discontinuities have disappeared.

If the rates are large, many folding operations are necessary, but the results are needed for one n interval only, whose lower and upper bounds $n_{\min}, n_{\max}$ have to be chosen to reach a desired accuracy. To speed up the numerical computations, it is advantageous to double the event numbers in every folding step until the minimal value of n is reached, and to keep the distributions for $n = 2^m$ with integer m for subsequent use. It is not necessary to compute folding integrals for all n. Distributions in the high n region can be computed partly by interpolation because the shapes are relatively stable. It is also possible to combine two histogram bins into one, if the number of X bins per standard deviation exceeds a cut with increasing n. This process can be iterated.

Finally the Poisson distribution for the appearance of n events has to be taken into account. If $\overline{n}$ is the mean rate, the final probability density is

$$P(X) = \exp(-\overline{n}) \cdot \delta(X) + \sum_{n \geq X/\max(w)}^{\infty} \exp(-\overline{n}) \cdot \frac{\overline{n}^n}{n!} \cdot P_n(X) \; . \tag{3.3}$$

For a given $X > 0$, only the terms with $n \geq X/\max(w)$ contribute.

Formula (3.3) is used to compute the complete distribution function $P_b(X)$ for background events. The distribution for signal events can also be obtained, and the result $P_s(X)$ has to be folded with $P_b(X)$ to obtain the overall distribution for signal and background, $P_{sb}(X)$.

The repetition of many folding operations would be time consuming if the signal rate r had to be modified iteratively. Therefore, the P_n distributions for fixed numbers of signal events, called P_{sn}, were folded with the complete background distribution $P_b(X)$ from (3.3). To compute confidence levels, only the cumulated distributions are needed:

$$C_n(X) = \int_0^X dZ \cdot \int_0^{\min(Z, n\cdot\max(w))} dY \cdot P_b(Z - Y) P_{sn}(Y) \; . \tag{3.4}$$

The cumulated distribution for the sum of signal and background is then

$$CL_{sb}(X) = \int_0^X P_{sb}(Y) dY = \exp(-r) \cdot \left(\exp\left(-\sum_{ki} b_{ki} \right) + \sum_{n=n_{\min}}^{n=n_{\max}} \frac{r^n}{n!} \cdot C_n(X) \right) \; . \tag{3.5}$$

These results can now be used to compute the expected confidence levels (2.7) and (2.8), needed to tune the R parameter:

$$\begin{aligned} E[CL_{sb}]_b &= CL_{sb}(X_{\text{cut}}) \quad \text{with} \quad CL_b(X_{\text{cut}}) = CL \; ; \\ E[CL_b]_{sb} &= CL_b(X_{\text{cut}}) \quad \text{with} \quad CL_{sb}(X_{\text{cut}}) = CL \; . \end{aligned}$$

The parameter CL replaces the parameter K in criteria (iv),(v) and (vii) and X_{cut} has to be computed from it by inversion of (2.1).

The expectation values needed for criteria (i) and (ii) are

$$< CL_{sb} >_b = \int_0^\infty CL_{sb}(X) P_b(X) dX \ ;$$
$$< CL_b >_{sb} = \int_0^\infty CL_b(X) P_{sb}(X) dX \ .$$

As already shown, both expectation values have their optimum at the same value of R.

A different numerical procedure to compute the series of folding integrals (3.5), based on a Fourier transformation, is given in ref. [21].

3.2 Some analytic results

The functions $P_1(w)$ and their statistical moments can be given analytically for a few ξ distributions, if the weight is proportional to the signal-to-background-ratio. According to equation (2.9) this corresponds to the limit $R \to 0$, which implies either a small signal-to-background-ratio or the lowest probability for a background fluctuation up to the median signal plus background level (criterion (iii)). Three cases will be discussed below.

- Gaussian signal, constant background. The w distribution for a Gaussian function $D(\xi) \sim \exp(-(\xi - \xi_0)^2/(2\sigma_\xi^2))$, its mean value, and its mean square for one signal event, are

$$P_{s1}(w) = \frac{1}{\sqrt{-\pi \cdot \ln w}} \ ; \quad < w >_s = \frac{1}{\sqrt{2}} \ ; \quad < w^2 >_s = \frac{1}{\sqrt{3}} \ . \tag{3.6}$$

  At the signal maximum the weight is set to 1. The background events are distributed according to

$$P_{b1}(w) = \mathcal{N} \cdot \frac{\sqrt{2}\sigma_\xi \frac{dB}{d\xi}}{w \cdot \sqrt{-\ln w}} \ . \tag{3.7}$$

  This equation contains a normalization factor $\mathcal{N}$, and with $\mathcal{N} = 1$ it gives the total background rate per w interval. The constant $dB/d\xi$ is the differential background rate. The expression is not integrable at $w = 0$ because an infinite number of events is taken into account far away from the signal. After truncation of the ξ spectrum the integral converges. The total mean and variance of w are finite even without the cutoff.

- Breit Wigner resonance, constant background. The convention here is

$$D(\xi) \sim \frac{1}{(\xi - \xi_0)^2 + \gamma^2} \ .$$

  The distribution, the mean and mean square of w for one signal event are

$$P_{s1}(w) = \frac{1}{\pi \cdot \sqrt{w \cdot (1 - w)}} \ ; \quad < w >_s = \frac{1}{2} \ ; \quad < w^2 >_s = \frac{3}{8} \ ;$$

  and the background distribution is

$$P_{b1}(w) = \mathcal{N} \cdot \frac{\gamma \frac{dB}{d\xi}}{w \cdot \sqrt{w \cdot (1 - w)}} \ .$$

- Two-dimensional Gaussian signal, constant background. Two independent discriminating variables are distributed according to $D(\xi,\eta) \sim \exp(-(\xi-\xi_0)^2/(2\sigma_\xi^2)) \cdot \exp(-(\eta-\eta_0)^2/(2\sigma_\eta^2))$. Instead of equations (3.6) and (3.7) one has

$$P_{s1}(w) = 1 \ ; \quad < w >_s = \frac{1}{2} \ ; \quad < w^2 >_s = \frac{1}{3} \ ;$$
$$P_{b1}(w) = \mathcal{N} \cdot \frac{2\pi\sigma_\xi\sigma_\eta}{w} \cdot \frac{\partial^2 B}{\partial\xi\partial\eta} \ .$$

From this the parameters needed for the high rate estimates of confidence levels in section 2.2 are obtained as

- Gaussian signal

$$< X >_s = \frac{r}{\sqrt{2}} \ ; \quad < X >_b = \sqrt{2\pi}\sigma_\xi\frac{dB}{d\xi} \ ; \quad \sigma_s^2 = \frac{r}{\sqrt{3}} \ ; \quad \sigma_b^2 = \sqrt{\pi}\sigma_\xi\frac{dB}{d\xi} \ .$$

- Breit Wigner signal

$$< X >_s = \frac{r}{2} \ ; \quad < X >_b = \pi\gamma\frac{dB}{d\xi} \ ; \quad \sigma_s^2 = \frac{3}{8}r \quad \sigma_b^2 = \frac{1}{2}\pi\gamma\frac{dB}{d\xi} \ .$$

- Two-dimensional Gaussian

$$< X >_s = \frac{r}{2} \ ; \quad < X >_b = 2\pi\sigma_\xi\sigma_\eta\frac{\partial^2 B}{\partial\xi\partial\eta} \ ; \quad \sigma_s^2 = \frac{r}{3} \ ; \quad \sigma_b^2 = \pi\sigma_\xi\sigma_\eta\frac{\partial^2 B}{\partial\xi\partial\eta} \ .$$

4. Applications

4.1 Upper limits without background subtraction

If nothing is known about the magnitude and the spectral shape of the background, upper limits for a signal rate can still be obtained by omitting the background in (3.5). The function CL_{sb} has then to be replaced by

$$CL_s(X) = \int_0^X P_s(Y)dY = \exp(-r) \cdot \left(1 + \sum_{n=n_{\min}}^{n=n_{\max}} \frac{r^n}{n!} \cdot \int_0^{\min(X,n\cdot\max(w))} dY \cdot P_{sn}(Y)\right) . \quad (4.1)$$

Apart from trivial event counting the only meaningful ansatz for the weights, valid for one search channel, is

$$w_{ki} = \frac{s_{ki}}{\max(s_{ki})} \ . \quad (4.2)$$

The upper rate limit, for a given CL_s, is obtained by solving (4.1) for r. The 95% exclusion limits ($CL_s = 0.05$) for Gaussian and Breit-Wigner ξ distributions are shown as a function of the test statistic in figure 3. For comparison, the figure also contains the 95% confidence limits from Poisson statistics without spectral sensitivity. In this case, the abscissa values are the observed event numbers.

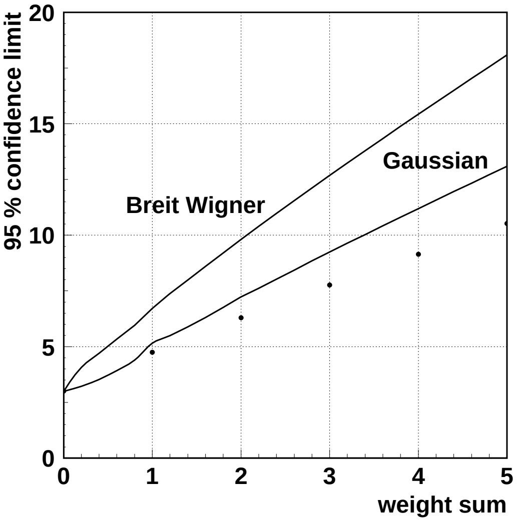

Figure 3: Excluded count rates with 95% confidence without background subtraction. lower curve: Gaussian distribution, upper curve: Breit-Wigner resonance. The dots at integer abscissa values are the Poissonian limits from unweighted counting.

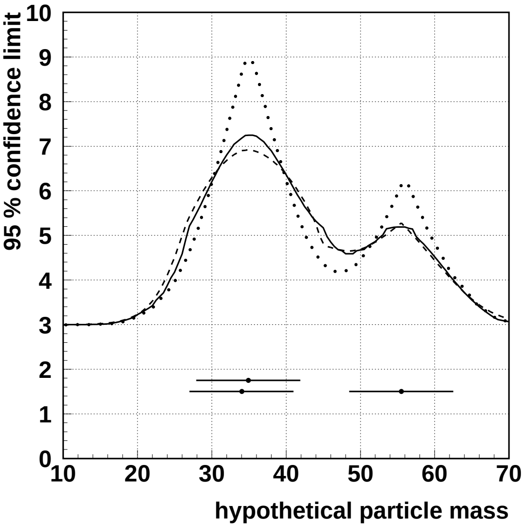

Figure 4: Limits on signal production rates from 3 events without subtraction of background. A Gaussian mass spectrum is assumed. The candidate positions are given by the points and the mass resolution is indicated by the error bars. Full curve: this work, dashed curve: Grivaz and Diberder, dotted curve: Gross and Yepes.

Figure 4 shows a 95% signal exclusion plot computed from three observed events, using their measured masses and varying a hypothetical resonance mass. Accidentally, two of the mass values are almost identical. The data set has been taken from ref. [7]. The mass resolution is assumed to be Gaussian. The results obtained with equation (4.1) are given by the solid line. The curve has kinks at the rate limit 5.2. This effect is visible in figure 3, too. It is due to the singularity of the distribution $P_1(w)$ at $w = 1$ (see figure 2). At the positions of the candidates, the rate limits are slightly worse compared to the Poissonian limits of 4.74 for one and 6.30 for two observed events. These more pessimistic results from fractional counting arise from theoretical configurations containing more events than the data sample but the test statistic being smaller than the observed one. This is the price one has to pay for mass discrimination. In mass regions away from the observed candidates, the rate limits from fractional counting are more stringent than the Poissonian limits.

The problem of obtaining mass selective rate limits without background subtraction has been discussed previously. Gross and Yepes [16] use fractional event counting, too. Their weight is defined as the probability that an arbitrary event has a larger mass difference with respect to the hypothetical particle than the candidate. In ref. [16] the incorrect assumption is made that the confidence limit for an integer number of fractional counts is equal to the Poissonian limit. The exclusions are too stringent. Nevertheless, the ansatz for the weight is a legitimate alternative, and the rate limits obtained with it, using the folding procedure (4.1), are added in figure 4. A disadvantage of this algorithm is that it produces unnatural sharp spikes at the candidate masses, and the limits at these positions lie far above the Poissonian limits.

Another formalism was given by Grivaz and Diberder [17]. They use a formula like the sum (4.1), truncated at the number of observed events, and the integrals are replaced by the probabilities that an arbitrary mass configuration of n events is less likely than the configuration of the n observed events closest to the hypothetical mass. The algorithm does no independent event counting. It has therefore the technical complication that corrections have to be applied to the equivalent of equation (4.1) to obtain the final unbiased probabilities [17]. Numerical results are also included in figure 4. They are very similar to those of this work.

4.2 Upper limits with background subtraction

If the background is known without any systematic error, a rate limit corresponding to a confidence level CL can be determined from the condition

$$CL = CL_{sb}(X_{obs}) \; . \tag{4.3}$$

The r dependence is given by (3.5), which contains the Poisson distribution. If the observed test statistic X_{obs} is smaller than the expectation from background it can happen that equation (4.3) has no positive solution. To avoid this problem the criterion to compute r is modified to [18, 6, 7]:

- The probability to observe a test statistic X smaller than or equal to the measured value X_{obs}, if the background contribution alone is $\leq X_{obs}$, is required to be smaller than or equal to CL.

This ansatz is motivated by the Bayesian treatment of background subtraction in counting experiments [19] and it gives an over-coverage by definition. The equation

$$CL = CL_s(X_{obs}) \qquad \text{with} \qquad CL_s(X_{obs}) = \frac{CL_{sb}(X_{obs})}{CL_b(X_{obs})} \; . \tag{4.4}$$

has to be solved for r.

Alternative procedures have been suggested which avoid the over-coverage as much as possible. The unified approach of Cousins and Feldman gives confidence belts instead of one-sided limits and has been applied to the Poisson and the Gaussian distribution [22]. The results for r are more stringent than those of (4.4). Algorithms with optimized coverage for the Bayesian procedure have been investigated by Roe and Woodroofe [23]. For the Poisson case this method can be shown to be related to equation (4.4) [23]. Other approaches with improved coverage are the ordering scheme of Punzi [24] and the profile likelihood method [25], where systematic errors have been included by Rolke, Lopez and Conrad [26].

The reasons for adopting (4.4) in this paper are the robustness of upper limits and the inclusion of the systematic errors in the event weight definition as described in section 6. It is a difficult and so far unsolved problem how to improve the coverage properties and to introduce this weighting scheme at the same time.

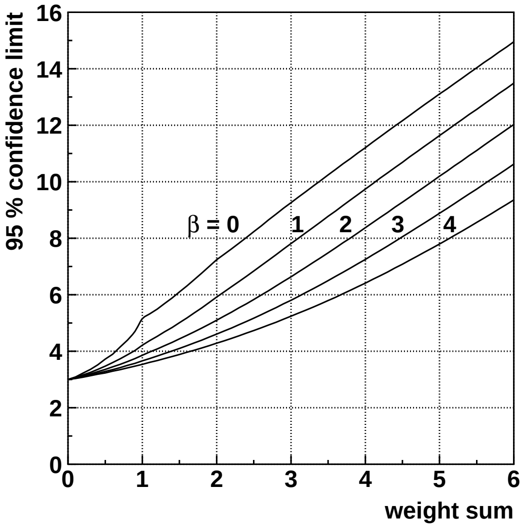

Figure 5: Count rates excluded with 95% confidence as function of the weight sum. The background is subtracted. The limits are for a Gaussian signal distribution and a constant background level β. The weight is taken proportional to the signal-to-background-ratio.

Figure 5 shows the 95% exclusion limits on r ($CL = 0.05$) for a Gaussian signal and constant background. To obtain the results the weight definition (4.2) was used, which is equivalent to $R \to 0$ in equation (2.9). The differential background level was varied and parameterized by its mean contribution to the test statistic $\beta =< X >_b= \sqrt{2\pi}\sigma_\xi \frac{dB}{d\xi}$. Asymptotically, the splitting between the curves in figure 5 becomes constant at large X. The rate limits computed with equation 4.4 are then lower than the results without background subtraction by an amount $\frac{\beta}{<w>_s} = \sqrt{2}\beta$.

According to section 2.2, the signal-to-background-ratio is not the best choice for the filter. Figure 6 shows the optimization of the R parameter for one special background level. It is based on the median expected limit $E[n_{95}]_b$, the theoretical signal rate n_95 which corresponds to $1 - CL_s = 0.95$, if only background contributes to the analyzed events. It is assumed that the test statistics has the median value β. Only a very weak dependence of $E[n_{95}]_b$ on R, of the order of $\delta = (E[n_{95}]_b - \min(E[n_{95}]_b)/\min(E[n_{95}]_b) = 0.3\%$, can be seen. The limits from a real observation can vary by several per cent, however, depending on the ξ positions of the events, and in general they are a monotone function of R.

Figure 7 gives the median expected limits $E[n_{95}]_b$ as a function of β. The lower curve indicates the R parameters used to obtain these results. At large β, one has $R \approx 1/2 \cdot E[n_{95}]_b$. The difference to the above estimate $R \approx E[n_{95}]_b$ is due to the fact that finally the limit computation is based onto CL_s and not onto CL_{sb}. Below $\beta = 1$ Poisson fluctuations play a significant role. There are several local minima of $E[n_{95}]_b$ if R is varied, and no solution has an obvious advantage over the others. The R values given in figure 7 below $\beta = 2$ are downward extrapolations consistent with $R = 0.4 \cdot E[n_{95}]_b$, which is the

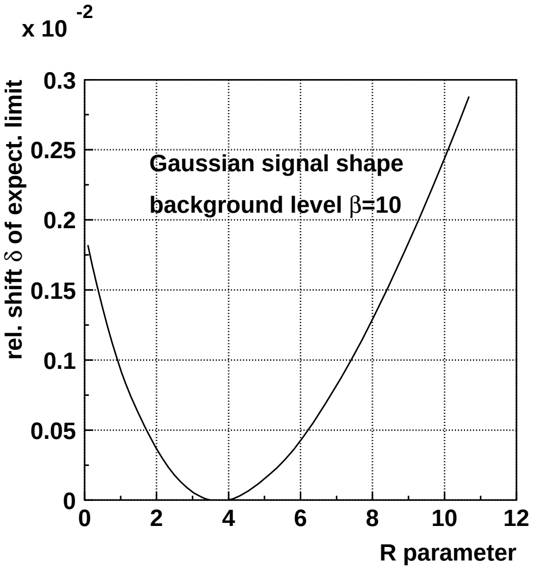

Figure 6: Dependence of the median expected 95% confidence limit on the rate parameter R. The background level is $\beta = 10$.

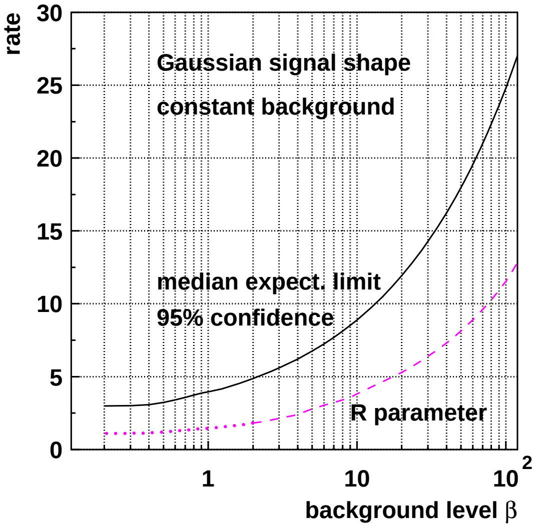

Figure 7: Median expected rate limits $E[n_95]_b$ as a function of the background level β, if no signal exists. The lower curve gives the parameters R used to compute the limits.

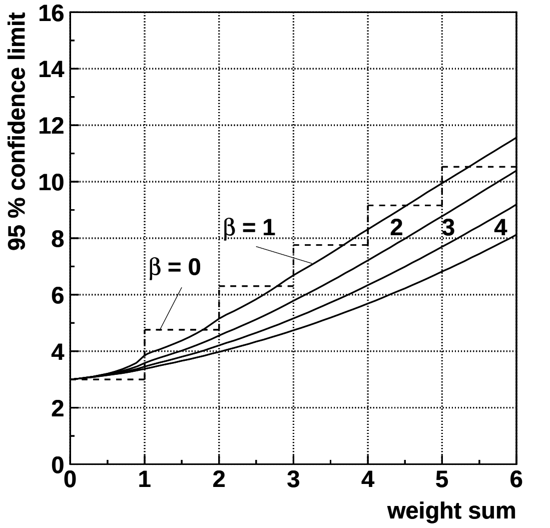

Figure 8: Count rates excluded with 95% confidence as function of the weight sum. The background is subtracted. The limits are for a Gaussian signal distribution and constant background levels β. The R parameters of figure 6 were used to define the weights (see text).

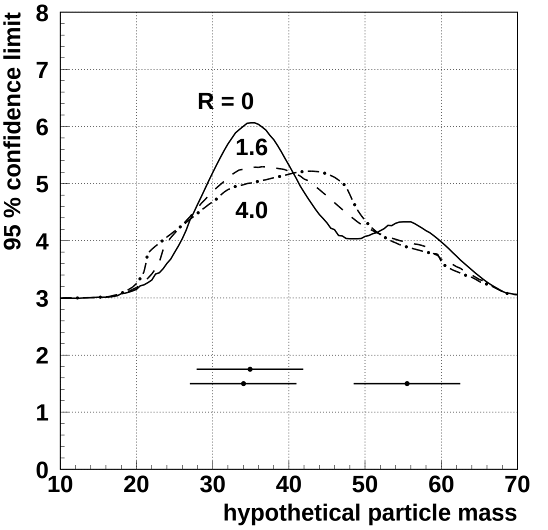

Figure 9: Limits on the production rate from 3 observed events with subtraction of 3 background events. The data are identical to figure 4. The curves are for different definitions of the weight algorithm and demonstrate the ambiguities in the analysis.

result around $\beta = 2$. They can be considered as upper bounds. Figure 8 is analogous to figure 5, but now the optimized R values from figure 7 are used in the analysis. It should be noted that the definition of X is not the same in figures 5 and 8, and in the latter case it is β dependent. The dashed step curve for $\beta = 0$ corresponds to the Poisson distribution because for any finite R and $\beta = 0$ the algorithm does normal event counting.

For small finite β the results depend strongly on R. This dependence is illustrated in figure 9. The example of figure 4 with three measured particle masses was analyzed again. This time it was assumed that a background of three events is predicted within the mass region of the plot, and this background was subtracted. It corresponds to $\beta = 0.88$. The limits are shown for three different weight definitions, all leading to legitimate results.

The parameter $R = 4$, resulting in a broad exclusion curve, is larger than the value from figure 7, which is approximately $R = 1.6$. The second exclusion curve in figure 9 corresponds to this value. The third curve for $R = 0$ corresponds, apart from background subtraction, to the result in figure 4. This ambiguity was already observed for the likelihood ratio method [7], the results of which are close to the case $R = 1.6$ shown here. Since the expected limit $E[n_{95}]_b$ depends only weakly on R it is recommended to keep the maximal mass resolution and to use $R = 0$ in low statistics experiments to resolve the above ambiguity.

5. Systematic errors

5.1 Parameterization of systematic errors

The errors are classified according to sources j. In principle every source may influence the ξ spectra of signal (s) and background (b) in all channels. It is parameterized by error functions $\sigma^{(s)}_{j,ki}$ and $\sigma^{(b)}_{j,ki}$ whose absolute values are the rms errors, given for channel k and bin i.

For the technical handling the following requirements were adopted in ref. [15]:

- Errors from the same source are treated as fully correlated between different bins of a signal or background histogram. The signs of the error functions determine the signs of the correlations.

- Errors from the same source are treated as fully correlated between signal and background.

- Errors from the same source are treated as fully correlated between different search channels.

- Errors from different sources are treated as uncorrelated. The spectra s_{ki} and b_{ki} are often available in an analytic form depending on parameters with correlated systematic errors. These correlations can be removed by diagonalizing the error matrix.

- The total relative error is much smaller than 100%.

The last requirement is not always satisfied. The error due to a mass resolution, for instance, has the same order of magnitude as the spectrum itself in bins far away from the mass peak and its distribution becomes asymmetric. However, it will be shown later in section 6 that such bins may be dropped.

The effect of systematic errors on confidence levels is commonly studied by Monte Carlo simulations [7]. To this aim the input spectra are modified according to

$$\begin{aligned} s^*_{ki} &= s_{ki} + \sum_j \sigma^{(s)}_{j,ki}\zeta_j \, , \\ b^*_{ki} &= b_{ki} + \sum_j \sigma^{(b)}_{j,ki}\zeta_j \end{aligned} \tag{5.1}$$

where the ζ_j are random numbers with mean zero and variance unity.

Often error functions $\sigma^{(s)}_{j,ki}, \sigma^{(b)}_{j,ki}$ related to likelihood or neural network variables are not well known if known at all. Usually, systematic errors are evaluated by modifying Monte Carlo simulations and counting the rate changes above an effective selection cut. The assumption is then made that the systematic errors have the same dependence on the discriminating variable as the signal or background distributions:

$$\begin{aligned} \sigma^{(s)}_{j,ki} &= \delta^{(s)}_{jk}\, s_{ki} \; ; \\ \sigma^{(b)}_{j,ki} &= \delta^{(b)}_{jk}\, b_{ki} \; . \end{aligned} \tag{5.2}$$

The remaining relative errors $\delta^{(s)}_{jk}, \delta^{(b)}_{jk}$ are still source and channel dependent but the bin dependence is ignored. In general, this ansatz is not valid and dictated by lack of knowledge. Nevertheless it has been applied in search experiments (for the Higgs boson, see ref. [15] and references therein). Thus this ansatz is used here as well.

5.2 Correction of confidence levels in the frequentist approach

In this subsection it is assumed that the sum X is a continuous variable. The distribution functions $P_{sb}(X)$ and $P_b(X)$ are not allowed to have delta-function-like singularities.

In the following considerations, the event source and the analysis hypothesis are the same: background events are analyzed in terms of background, and the same is done for the combination of signal and background. The indices of the functions CL_{sb}, CL_b are dropped for simplicity.

For the case of a correct analysis hypothesis with the correct production rate the CL values are uniformly distributed between zero and one. An ensemble of observers is introduced to average over the modified parameter sets given by equation (5.1). The distribution function $D_{\text{rec}}(CL_{\text{rec}})$ of the reconstructed confidence levels CL_{rec}, as evaluated by one observer, is not constant. The modifications (5.1) introduce variable slopes. Because CL_{rec} has a lower and an upper bound, the function $D_{\text{rec}}(CL_{\text{rec}})$, averaged over the ensemble, peaks at 0 and 1. Without any correction, the observers reconstruct too often data deficits or excesses, as illustrated in figure 10.

A correction can be made in the following way. The distribution D_{rec} has to be integrated up to the reconstructed confidence level CL_{obs} of a certain observer and the integral obtained is then the corrected confidence level (see figure 10):

$$CL_{\mathrm{corr}} = \int_0^{CL_{obs}} D_{\mathrm{rec}}(CL_{\mathrm{rec}}) dCL_{\mathrm{rec}} \ . \tag{5.3}$$

This equation, known as the probability integral transform, is valid for arbitrary shapes of the ζ_j distributions.

The procedure has to be applied independently to both CL_b and CL_{sb}. The problem is that one observer of the ensemble cannot reconstruct the function D_{rec} of the ensemble because the true physical parameters are unknown. The observer necessarily takes his own spectra $s(\xi)$, $b(\xi)$, instead of the true ones, to evaluate the correction of CL_{obs}. This replacement is unavoidable and causes deviations of the average CL_{corr} distribution from uniformity.

Equation (5.3) is not convenient for numerical calculations. Instead, CL_{corr} should be expressed as an integral over the stochastic variables ζ_j.

As already mentioned, an observer starts with original signal and background spectra s_{ki}, b_{ki}, from which the test statistics X_o is constructed. The mean value and the rms error of X_o are denoted by $< X_o >$ and σ_o, respectively. The signal and background distributions are modified according to equations (5.1). The new spectra are used to redefine the event weights. The original spectra can then be inserted into equations (2.3) with the new weights used. The mean value of the test statistic $< X_o >$ and its rms error σ_o are shifted to $< X >$ and σ. The complete folding according to section 3.1 gives the function $CL_{\mathrm{orig}}(X, \vec{\zeta})$. For clarity the ζ_j-dependence of CL_{orig} is stated explicitly here. The modified spectra have to be analyzed, too, resulting in the distribution $P_{\mathrm{rec}}(X, \vec{\zeta})$ with the statistical parameters $< X^* >$, σ^* and its integral $CL_{\mathrm{rec}}(X) = \int_{-\infty}^{X} P_{\mathrm{rec}}(Y, \vec{\zeta}) \cdot dY$. To get the integration limit in equation (5.3) and figure 10 the last integral has to be identified with CL_{obs}

$$CL_{obs} = \int_{-\infty}^{X^*(\vec{\zeta})} P_{\mathrm{rec}}(Y, \vec{\zeta}) dY \ . \tag{5.4}$$

The probability for the test statistic X of the original distribution to be lower than or equal to the upper limit $X^*(\vec{\zeta})$ is $CL_{\mathrm{orig}}(X^*, \vec{\zeta})$. To get the integral (5.3) a summation is performed over all Monte Carlo experiments. This leads to the final result

$$CL_{\mathrm{corr}}(CL_{obs}) = \int_{-\infty}^{\infty} \prod_j d\zeta_j \cdot P_{\mathrm{sys}}^{(f)}(\vec{\zeta}) \cdot CL_{\mathrm{orig}}(X^*, \vec{\zeta}) \ , \tag{5.5}$$

where $P_{\mathrm{sys}}^{(f)}(\vec{\zeta})$ is the distribution of the random vector $\vec{\zeta}$ in the frequentist approach and the argument X^* is taken from the condition (5.4).

5.3 Equivalence between the frequentist and the Bayesian treatment of systematic errors

Usually systematic errors are treated with the Bayesian method introduced by Cousins and Highland [12]. Their formula for the Poissonian case can be generalized for the situation with event discriminators in many channels. For systematic errors with no correlations between bins the multi-channel-case has been treated by Lista [27].

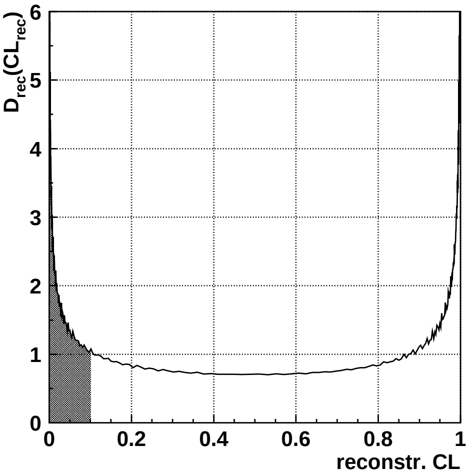

Figure 10: Distribution of reconstructed confidence levels for a Poisson distribution with a mean rate of 100 events and a systematic error of 10%, as reconstructed by many observers.

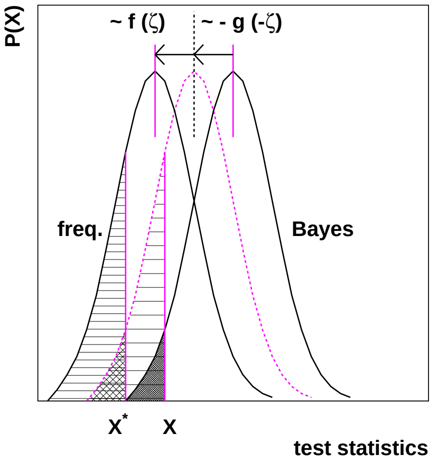

Figure 11: Relationship between the frequentist and the Bayesian treatment of systematic errors. Central curve: original distribution of the test statistic. X is a measured value. Right curve: shifted distribution according to Bayesian error treatment for $\zeta < 0$. Dark area: contribution to the corrected confidence level. Left curve: shifted distribution according to the frequentist approach for the same value of ζ. The two horizontally hatched areas are equal by construction. The agreement of both approaches is guaranteed if the small marked areas are equal.

As mentioned above, an observer does not know the true ξ spectra but only the estimates s_{ki} and b_{ki}. The possible variants of the true spectra can be described by a set $\vec{\zeta_B}$ of stochastic variables. Again a function $P_{\rm sys}^{(B)}(\vec{\zeta_B})$ is introduced, the probability that the set $\vec{\zeta_B}$ is the correct one. The reconstructed confidence levels depend on $\vec{\zeta_B}$. Now two different statistical methods are be mixed: to include systematic errors, the confidence levels from the frequentist approach are folded with the observers prejudice on the true spectra:

$$CL_{\text{corrected}} = \int_{-\infty}^{+\infty} \prod_j d\zeta_{Bj} \cdot P_{\text{sys}}^{(B)}(\vec{\zeta_B}) \cdot CL_{\text{rec}}(X, \vec{\zeta_B}) \ . \tag{5.6}$$

Here, X is the measurement of the observer.

The theoretical spectra enter the analysis twice: firstly they are needed to compute the weight function, and secondly the absolute rates are used in the statistical analysis of section 3.1. It has always been a matter of debate whether systematic errors should be assigned to the weight definition w_{ki}, and it has become practice to keep this function fixed [11, 15]. The argument for this is that the weight definition is arbitrary and the results are correct for any fixed definition. Within the frequentist approach, this argument is not correct for a principle reason: it is highly unlikely that two observers use exactly the same numerical parameters for their data analysis. Every observer constructs his own weight function, and the only possible agreement is a common value of the ratio R/r relevant for equation (2.9). Nevertheless, in the following a fixed weight definition is adopted because it is then easy to compare the frequentist and the Bayesian approaches. One has therefore $X_o = X$ and $< X_o > = < X >$.

Equations (5.6) and (5.5) look completely different. However, a wide class of X densities, for which both approaches agree, can be constructed assuming shape invariance of the X distribution:

$$\begin{aligned} P_{\text{orig}}(X) &= P_{\text{rec}}(X, \vec{\zeta} = 0) = F\left(\frac{X - < X >}{\sigma}\right) , \\ P_{\text{rec}}(X, \vec{\zeta}) &= F\left(\frac{X - < X^* >}{\sigma^*}\right) \ . \end{aligned} \tag{5.7}$$

The denominators are the rms errors of the test statistic and F is a universal function independent of the stochastic variables ζ_j. One has to distinguish the shifted parameters $< X^* >, \sigma^*$ for the frequentist case from those of the Bayesian treatment, indicated by superscripts (f) and (B). A constant reconstructed confidence level means that the ratio $(X - < X^{*(f)} >)/\sigma^{*(f)}$ is the same for $\vec{\zeta} \neq 0$ and $\vec{\zeta} = 0$:

$$\begin{aligned} \frac{X^* - < X^{*(f)} >}{\sigma^{*(f)}} &= \frac{X - < X >}{\sigma} \\ X^* = X \cdot \frac{\sigma^{*(f)}}{\sigma} &+ < X^{*(f)} > - < X > \cdot \frac{\sigma^{*(f)}}{\sigma} \ . \end{aligned} \tag{5.8}$$

The ansatz for the systematic error within the frequentist approach is

$$\frac{< X^{*(f)} >}{\sigma^{*(f)}} = \frac{< X >}{\sigma} + f(\vec{\zeta}) \cdot \frac{\sigma_{\text{sys}}^{(f)}}{\sigma} \ . \tag{5.9}$$

The arbitrary function f of the stochastic variables $\vec{\zeta}$ describes non-Gaussian systematic errors. Its variance is normalized to unity. Equivalently, for the Bayesian treatment the parameterization is given by

$$\frac{< X^{*(B)} >}{\sigma^{*(B)}} = \frac{< X >}{\sigma} + g(\vec{\zeta_B}) \cdot \frac{\sigma_{\text{sys}}^{(B)}}{\sigma} \ . \tag{5.10}$$

A set of Bayesian stochastic variables $\vec{\zeta_B}$ is introduced here. The vector $\vec{\zeta}$ parameterizes a shift from the original X distribution to the function used by an arbitrary observer. In the Bayesian interpretation, the direction of the shift has to be inverted, so that $\vec{\zeta_B} = -\vec{\zeta}$.

A condition sufficient for the equivalence of (5.6) and (5.5) is then

$$P_{\text{rec}}(X, \vec{\zeta_B}) = P_{\text{rec}}(X, -\vec{\zeta}) = P_{\text{orig}}(X^*) \cdot \frac{dX^*}{dX} \tag{5.11}$$

for any X and $\vec{\zeta}$.

Equations (5.7) to (5.10) have to be inserted into (5.11). The Bayesian quantities $< X^{*(B)} >, \sigma^{*(B)}$ enter the left hand side and the frequentist quantities $< X^{*(f)} >$ and $\sigma^{*(f)}$ the right hand side. A relationship between the widths $\sigma^{*(B)}, \sigma^{*(f)}$, the systematic errors $\sigma_{\text{sys}}^{(f)}, \sigma_{\text{sys}}^{(B)}$, and also the functions f and g has to be found. Consistency in equation (5.11), and thus equivalence between the two approaches, can be obtained with

$$\begin{aligned} \sigma_{\text{sys}}^{(f)} &= \sigma_{\text{sys}}^{(B)} \, , \\ \sigma^{*(B)}(-\vec{\zeta}) &= \sigma^{*(f)}(\vec{\zeta}) = \sigma \, , \\ -g(-\vec{\zeta}) &= f(\vec{\zeta}) \; . \end{aligned}$$

In general, this derivation of the equivalence fails if one introduces an arbitrary $< X >$ dependence into σ^*, σ_{sys} or f and g, or if the shape invariance (5.7) is violated. The meaning of the symmetry requirement on f, g is illustrated in figure 11.

The above arguments for the equivalence of both approaches have the following general features:

- The distribution of the test statistic may be an arbitrary continuous function.
- The distribution of systematic shifts may have an arbitrary shape. Even one-sided shifts are allowed.

To guarantee the equivalence between the frequentist and the Bayesian treatment of systematic errors, the combination of the following conditions is sufficient but not necessary:

- A fixed weight function is assumed, so that the test statistic for one experiment is independent of the observer.
- The systematic errors shift the X distributions but do not change their shapes.
- In addition, the $\vec{\zeta}$ distribution of systematic errors is invariant against translations of the X distributions.

In general, the Bayesian and the frequentist approach do not agree if one of the last two conditions is not fulfilled. If both criteria are violated, the effects can compensate each other by chance. This requires, however, an unnaturally fine tuning between the functions $P_{\text{rec}}(X, \vec{\zeta})$ and the error functions $f(\vec{\zeta})$ and $g(\vec{\zeta})$.

5.4 Numerical treatment of systematic errors

The numerical treatment here is limited to symmetric systematic errors. As a consequence, the confidence level shifts are proportional to the mean squares of errors if the latter are small. Asymmetric errors modify the expectation values $< X >_b$ and $< X >_{sb}$ in first order and have larger impacts. It is easy to include the Bayesian treatment of systematic errors in a computer program and this code can also be applied to low counting rates. A repetition of the folding operations (3.2) inside a Monte Carlo loop based on (5.1) is, however, very time consuming. A faster program can be set up using the shape invariance (5.7) and the additivity assumption (5.10), together with Gaussian distributions for the systematic errors of $< X >_b$ and $< X >_{sb}$.

The inclusion of systematic errors into the final results is then straightforward: With the help of N_{MC} Monte Carlo experiments (5.1) and the definitions (2.3) and (2.4) the systematic error is obtained from

$$\chi^2_{\mathrm{sys}} = \frac{1}{N_{MC}} \sum_{\mathrm{MC\ experiments}} \left(\frac{< X^* >}{\sigma^*} - \frac{< X >}{\sigma} \right)^2 .$$

In this expression an arbitrary scaling factor in the w_{ki} cancels and the expectation values involved can be computed without the folding procedure (3.2).

Equation (5.6), together with (5.7) and (5.10), leads to a folded distribution, from which the corrected confidence levels can be computed:

$$\begin{aligned} CL_{\mathrm{corr}}(X) &= \int_0^X dY \cdot P_{\mathrm{corr}}(Y) \qquad \text{with} \\ P_{\mathrm{corr}}(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\zeta \cdot \exp(-\zeta^2/2) \cdot P_{\mathrm{orig}}(X + \zeta\chi_{\mathrm{sys}}\sigma) \ . \end{aligned} \tag{5.12}$$

The parameter χ^2_{sys} is different for background alone and a combination of signal and background, and it depends on the overall signal-to-background-ratio. If the signal rate r is modified to find a rate limit, χ_{sys} has to be reevaluated.

The procedure has the advantage that it avoids a conceptual problem for the extraction of rate limits from CL_s. Without systematic errors, CL_{sb} is a monotonic function of CL_b if the test statistic is eliminated. Since this function becomes observer dependent in the presence of systematic errors, it is unclear how CL_s should be defined. In the above approach the ratio of folded functions CL_{sb} and CL_b is the natural choice. The method has been suggested by Zech [20] for counting experiments.

5.5 Poisson distribution at small rates

Even if the frequentist and Bayesian handling of systematic errors agree with each other, it is not guaranteed that the latter one is correct at low rates. In this region the Poisson distribution violates the criterion of shape stability as required in subsection 5.3. It was therefore investigated whether the Bayesian treatment gives a reasonable spectrum of reconstructed confidence levels, for the Poisson distribution. As an extreme case which has practical relevance for background estimates, the problem was studied for a very small

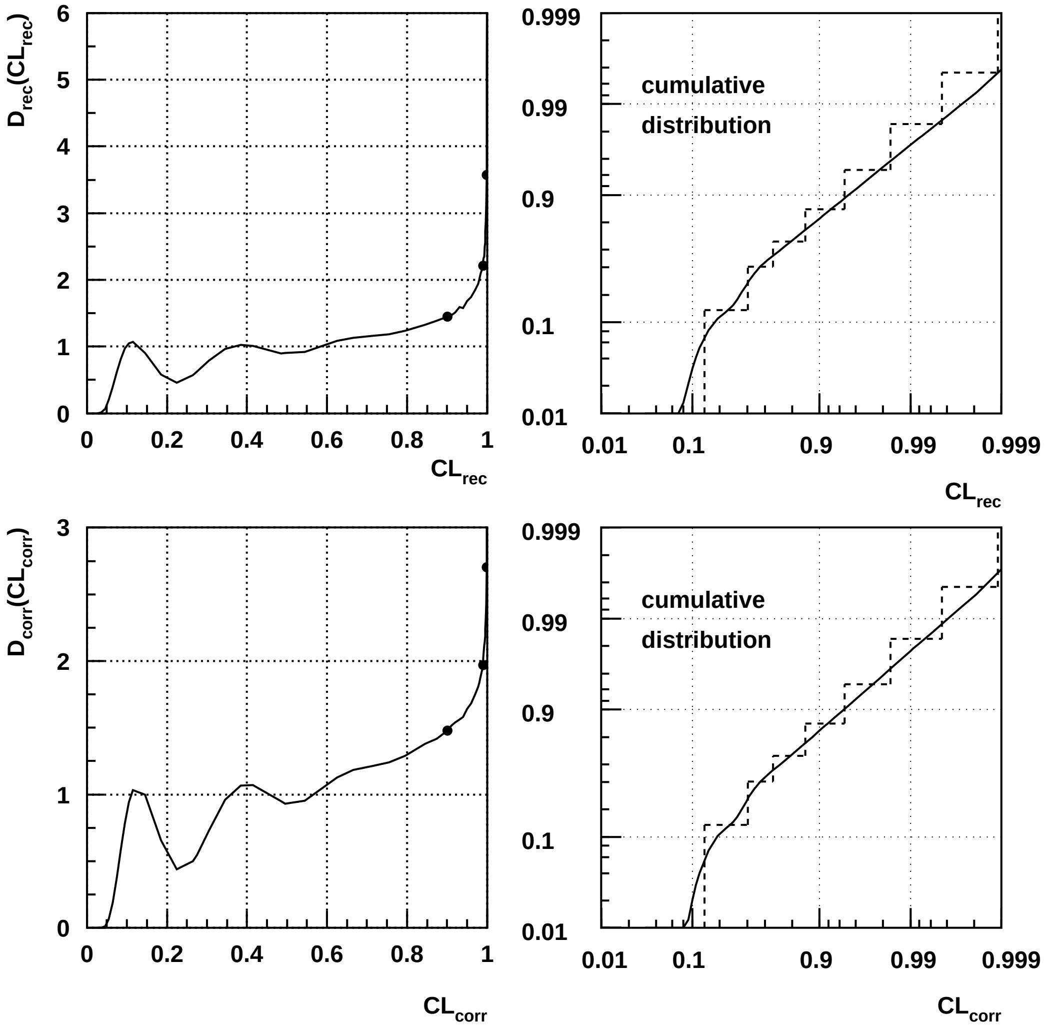

Figure 12: Spectra of confidence levels for a Poisson distribution with $\overline{n_0} = 2$ and a systematic error of 20%, as reconstructed by many observers. Lower (upper) part: The confidence levels are corrected (not corrected) for systematic errors. Left: differential distributions. The dots mark the results for the abscissa values 90%, 99% and 99.9%. Right: cumulative distributions. The step functions show the true original cumulative distribution of CL.

mean rate $\overline{n_0} = 2$ with a large Gaussian systematic error of 20%. The formalism how to get corrected confidence levels was taken from ref. [12].

The following test was made. An ensemble of observers was introduced with different choices for the mean rate. For any observer a new Poisson distribution was generated and, for any number of counts, n, the confidence levels $CL_{\mathrm{rec}}(n)$ and $CL_{\mathrm{corr}}(n)$ were computed, with the observer's mean rate n_0. Here, $CL_{\mathrm{rec}}(n)$ is defined without systematic errors and $CL_{\mathrm{corr}}(n)$ includes the correction from ref. [12]. The results were histogrammed with weights equal to the true probability to find n counts.

Figure 12 shows the resulting differential as well as the cumulative distributions. The differential spectrum of corrected confidence levels is found to be non-uniform at high CL_{corr}, where it still exhibits a spike at $CL = 1$. The right column of figure 12 shows the cumulative CL distribution in a special logarithmic representation. The result does not approach the diagonal at $CL = 1$. One could blame this on the fact that the same relative error was assumed for all observers and that a more adequate choice would be $\delta \sim 1/\sqrt{n_0}$. Tests have shown that this ansatz gives some improvement but does not cure the problem.

As an example, we assume that eight events are observed. To get the probability for a fluctuation from $\overline{n_0}$ to eight events or more one has to compute the confidence level CL for seven events. Poisson statistics without systematic errors give $1 - CL = 0.0011$, the corrected value according to ref. [12] is $1-CL = 0.0018$. With these numbers one obtains, from both graphs in the right column of figure 12, a probability of 0.0050 which should be quoted as a more realistic estimate, instead of the value 0.0018.

A clear conclusion is that indications for discoveries obtained from low statistics samples should be considered with great care if the background has a substantial uncertainty. Even after standard corrections for systematic errors the significance of the observation is still overestimated. For a given experiment this bias has to be investigated.

There is one exceptional case where the treatment of the systematic error in a low statistics experiment is correct. This example is the counting of the mean value $\overline{n_0}$ in a Monte Carlo simulation and can be found in ref. [28]. With a mathematical theorem given in that paper it can be shown that the Bayesian treatment of the statistical uncertainty of $\overline{n_0}$ is correct if the Monte Carlo and data luminosities are the same. Both the criteria of shape stability of the n distribution and the translational invariance of the systematic errors under shifts of $\overline{n_0}$ are violated here. In this very special situation both of these effects cancel in the evaluation of the systematic error.

6. Event weighting with systematic errors

In the preceding section systematic errors were included in the final results but the weight function (2.9) was optimized with respect to the statistical errors only. If search channels with very different systematic errors have to be combined or if many low weight background events contribute to fractional counting, this is not the best way to analyze the data. Instead, bins with large systematic errors should be downgraded in the analysis.

The procedure described in section 2.2 can be generalized to do this. Again the limiting case of Gaussian distributions for the test statistic is considered here. The generalization is given below for the criteria (i,ii),(iii) and (vi) of section 2.2.

The optimization criteria (i,ii) minimize the inverted ordering of the test statistic for arbitrary data sets corresponding to hypotheses (A) and (B). The supplementary condition that the total weights X for the comparison are measured by independent, arbitrary observers, has to be introduced.

All optimization criteria are the same as in section 2.2 except that the systematic errors are included in the variances σ_{sb}^2 and σ_b^2. Their contributions are obtained with equations (2.3) and (5.1):

$$
\begin{aligned}
\sigma_{sb}^2 &= \sum_{ki} w_{ki}^2 \cdot (s_{ki} + b_{ki}) + \sum_j \left(\frac{\partial < X >_{sb}}{\partial \zeta_j} \right)^2 = \\
&= \sum_{ki} w_{ki}^2 (s_{ki} + b_{ki}) + \sum_j \left(\sum_{ki} w_{ki} \left(\sigma_{j,ki}^{(s)} + \sigma_{j,ki}^{(b)} \right) \right)^2 ; \\
\sigma_b^2 &= \sum_{ki} w_{ki}^2 \cdot b_{ki} + \sum_j \left(\frac{\partial < X >_b}{\partial \zeta_j} \right)^2 = \\
&= \sum_{ki} w_{ki}^2 b_{ki} + \sum_j \left(\sum_{ki} w_{ki} \sigma_{j,ki}^{(b)} \right)^2 .
\end{aligned}
$$

The optimization, performed in analogy to section 2.2, leads to

$$
\begin{aligned}
& w_{ki} \cdot (s_{ki} k_1 + b_{ki} k_1 + b_{ki} k_2) + \sum_{lm} w_{lm} \cdot \sum_j \left(\sigma_{j,lm}^{(s)} + \sigma_{j,lm}^{(b)} \right) \left(\sigma_{j,ki}^{(s)} + \sigma_{j,ki}^{(b)} \right) \cdot k_1 + \\
& + \sum_{lm} w_{lm} \cdot \sum_j \sigma_{j,lm}^{(b)} \sigma_{j,ki}^{(b)} \cdot k_2 = s_{ki} . \qquad (6.1)
\end{aligned}
$$

The numerical factors k_1 and k_2 depend on the optimization criterion. One has $k_1 = k_2 = 1$ for conditions (i),(ii), $k_1 = 0$ for (iii) and $k_2 = 0$ for (vi). If the systematic errors are set to zero, equations (6.1) reduce to (2.9). The double sums correct the weights (2.9) for systematic errors, but they contain the final result w_{ki} so that the system of linear equations (6.1) has to be solved for w_{ki}.

Negative values are possible for the weights. Mathematically there is nothing wrong with this. The algorithm tries to extract information on the background from ξ bins with low signal content and to extrapolate it into the most significant signal region to improve the accuracy. However, when the approximation (5.2) is inserted into equation (6.1), the errors on the shapes of the ξ distributions are ignored and the appearance of negative weights is unacceptable. The problem can be cured if equations (6.1) are supplemented by the requirement that negative w_{ki} should not be allowed. It can be shown that the equations (6.1), together with these conditions, have a unique solution. This solution, with a reduced number of contributing bins, gives an improved discrimination between hypotheses (A) and (B) (see appendix A for a more detailed discussion).

This is illustrated in figure 13, which shows the expected upper rate limits for a Gaussian signal, computed for a constant background. On the left hand side of figure 13, the original weights (2.6) are compared with the result of (6.1). It turns out that the region of accepted events around the signal peak is rather narrow if the systematic errors are comparable to the statistical ones. The acceptance window depends on the background level β, which is again the number of events in the ξ interval $\sqrt{2\pi}\sigma_\xi$. The expected rate limits (see figures on the right hand side) with the filter (6.1) (full lines) are lower than the limits computed with (2.9) (dotted lines). It is also evident from the figure that the

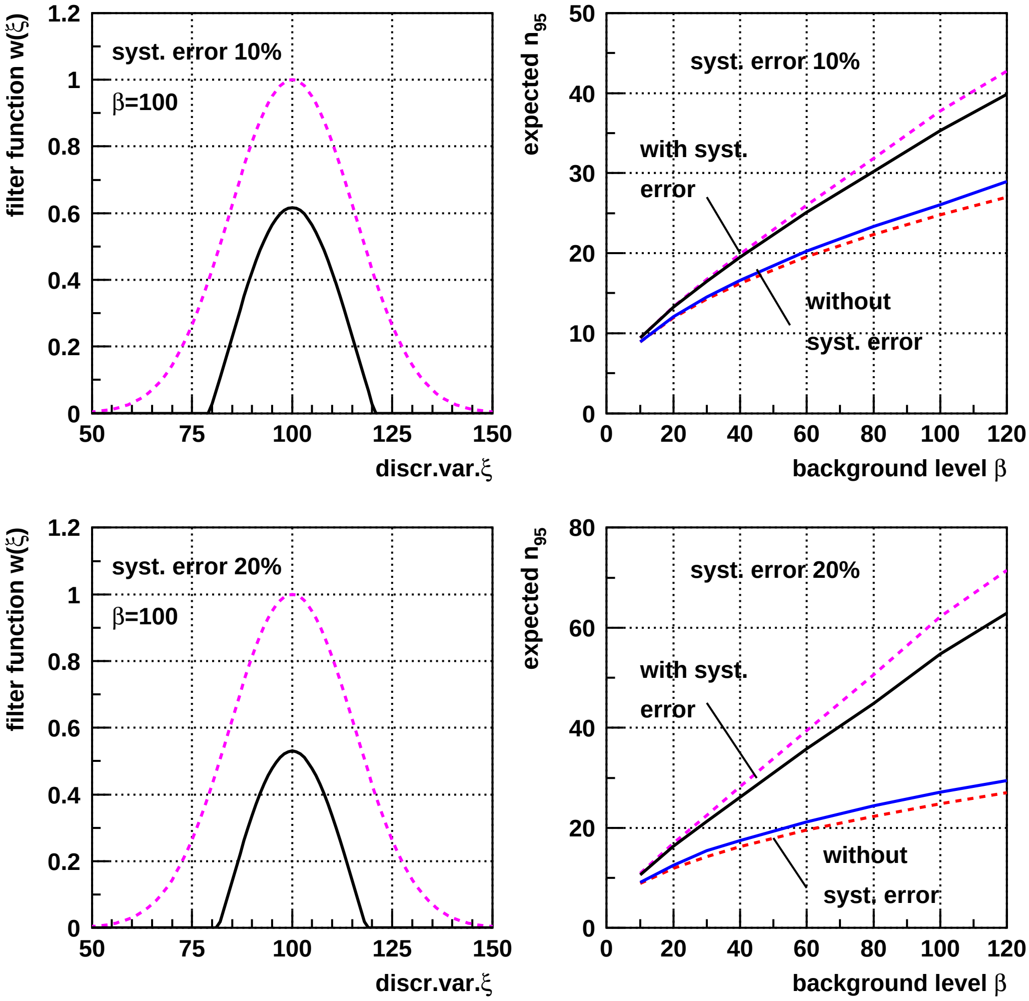

Figure 13: Expected upper rate limits $E[n_{95}]$ for a non-existing Gaussian signal over a constant background. Left column: weight functions. Dashed curves: weighting based on statistical errors only. Full curves: systematic errors included in the weights. The ordinate scale is arbitrary. Right column: 95% limits as a function of the background level. The line styles indicate the weight functions used in the analysis.

ordering of curves for the same weight functions is inverted if the systematic errors are not included in the statistical analysis, this is a consequence of the bin dropping.

Results of similar quality can be obtained with (2.9), together with a cut on s_{ki}/b_{ki}. This would, however, require the tuning of another parameter. Apart from the additional degree of freedom, this procedure would contradict our motivation for introducing fractional counting namely to avoid hard cuts in the event acceptance.

The weighting method discussed here should be applied if the systematic errors, including their correlations, have the same order of magnitude as the statistical errors or are even larger. A relevant physical example is the flavor-independent search for Higgs bosons [29]. Compared to more specific Higgs searches, the background here is larger and the relative systematic uncertainties are similar. Measured upper limits have an error component proportional to the background level so that systematic errors become important.

7. Confidence levels from the shapes of distributions

The methods presented in the preceding sections do not check at all whether the shapes of the underlying distributions ϵ_{ki} and b_{ki} are consistent with observation. A large value of of the measured sum X_{obs} normally indicating a discovery might also be due to an excessive number of background events. If the observed X_{obs} is close to unity, it is an interesting question to what extent this result is due to the number of events or to the difference in shape of the distributions $s(\xi)$ and $b(\xi)$.

A statistical test which does not compare the observed rate with a prediction but is sensitive to the local signal-to-background ratios only is proposed. The probability for an event, observed at ξ_{ki}, to be a signal event, is given by

$$p_{ki} = \frac{s_{ki}}{s_{ki} + b_{ki}} \ . \tag{7.1}$$

The predicted rates are taken from hypothesis (B). An arbitrary set of n_{obs} events obeys the polynomial distribution. From the observed candidates a likelihood $L^{(\mathrm{shape})}$ is calculated as

$$L^{(shape)} = \prod_{l=1}^{n_{obs}} p_{k(l)i(l)} \ .$$

A confidence level $CL_{sb}^{(shape)}$ is then defined as the probability that an arbitrary experiment with the same number of candidates gives at most the likelihood of the observed configuration. The same analysis can be made using the definition (7.1) based on hypothesis (B) but assuming that all events arise from background. The resulting confidence level is called $CL_b^{(shape)}$. Values of $CL_{sb}^{(shape)}$ or $CL_b^{(shape)}$ between 0.16 and 0.84 indicate consistency with the tested models, within one standard deviation. If CL_b is close to unity but $CL_{sb}^{(shape)}$ is small a discovery is ruled out. Vice versa, a $CL_b^{(shape)}$ value close to unity supports a discovery. If both $CL_b^{(shape)}$ and $CL_{sb}^{(shape)}$ are consistent with the underlying hypotheses, the test is not conclusive, either because the spectral shapes of signal and background are too similar or because the overall signal-to-background-ratio is too small.

A computation of the confidence levels requires the distribution functions of $L^{(shape)}$. The variable $L^{(shape)}$ can be replaced by its logarithm. The test corresponds then to fractional counting of a fixed number of events with the weight

$$w_{ki} = \ln \frac{s_{ki}}{s_{ki} + b_{ki}} \ . \tag{7.2}$$

The folding procedure is the same as in section 2 with the exception that the lower weight limit becomes negative. The algorithm has the same disadvantage as the likelihood ratio

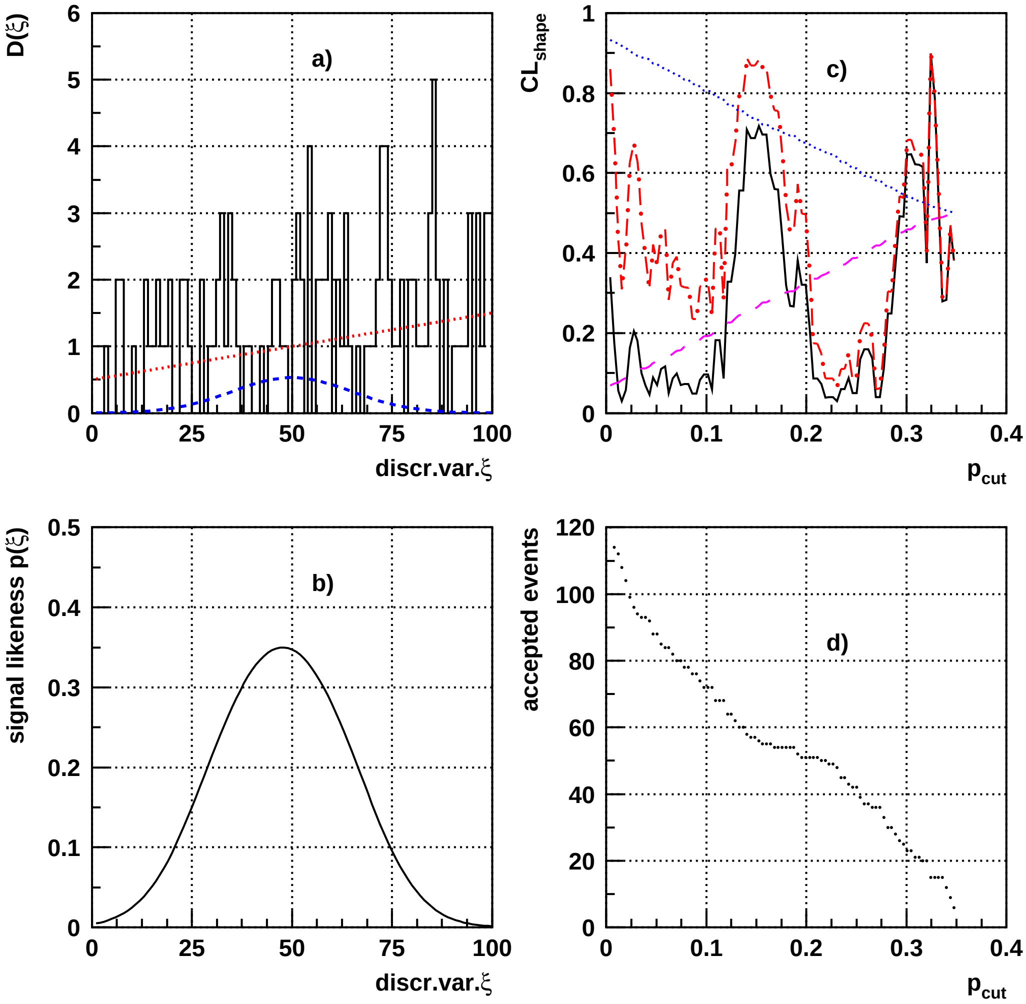

Figure 14: Confidence levels based on the polynomial distribution for a simple example. Upper left: signal (dashed), background (dotted) and candidate distributions (histogram). Lower left: the signal probabilities $p(\xi)$. Upper right: confidence levels. The full and the dash-dotted curves are for the 'data', the smooth curves are median expectations (see text). The analysis hypothesis is background for the upper curves, signal and background for the lower curves. Lower right: number of events with $p(\xi) \geq p_{\rm cut}$.

method namely a singularity, this time at s_{ki}=0. To avoid numerical problems, p_{ki} is required to exceed a minimum value. A continuous upward shift of this cut $p_{\rm cut}$ removes one candidate after the other from the sample, until the results are not anymore conclusive. The values of $CL_b^{(shape)}$ and $CL_{sb}^{(shape)}$ jump at the discontinuities.

As a simple example, figure 14a shows a Gaussian signal peak, a linearly rising background and a pattern of candidate events. The mean values are 100 (background) and 20 (signal), and the resolution is 15 bins. Compared to the background model, the sample contains an excess of events (130). The example experiment is analyzed at the hypothetical signal position. The normal analysis of section 4.2 gives confidence levels $CL_b = 0.993$ and $CL_{sb} = 0.45$, which would be a weak indication for a signal consistent with the assumed signal rate. Figure 14c shows the confidence levels $CL_b^{(shape)}$ and $CL_{sb}^{(shape)}$ as a function of p_{cut}. The smooth curves are two theoretical predictions: background events are analyzed in terms of signal plus background (lower curve) and signal plus background events are investigated assuming all events are background (upper curve). The number of accepted events as a function of p_{cut} is given, too. The falling sensitivity with decreasing number of events is obvious from the pictures.

In the sensitive region $p_{\mathrm{cut}} < 0.05$ there is a slight preference for a background interpretation, but the results are not very stable if p_{cut} is varied. This raises the question whether an analysis of this type can be performed without a cut on the weights. This is indeed possible by applying the procedure described in section 2.2. The variances of the ξ distributions for hypotheses (A) and (B) for one event are

$$\sigma_b^2 = \sum_{ki} \left(w_{ki} - \frac{\sum_{lj} b_{lj} w_{lj}}{\sum_{lj} b_{lj})} \right)^2 \cdot \frac{b_{ki}}{\sum_{lj} b_{lj}} \tag{7.3}$$

$$\sigma_{sb}^2 = \sum_{ki} \left(w_{ki} - \frac{\sum_{lj}(s_{lj} + b_{lj}) \cdot w_{lj}}{\sum_{lj}(s_{lj} + b_{lj}} \right)^2 \cdot \frac{s_{ki} + b_{ki}}{\sum_{lj}(s_{lj} + b_{lj})} \tag{7.4}$$

A possible normalization of w_{ki} is

$$\sum_{ki} \frac{(s_{ki} + b_{ki}) \cdot w_{ki}}{\sum_{lj}(s_{lj} + b_{lj})} - \sum_{ki} \frac{b_{ki} \cdot w_{ki}}{\sum_{lj} b_{lj}} = \text{const.} \tag{7.5}$$

In these equations the absolute normalization of the Monte Carlo rates s_{ki} and b_{ki} cancels. Instead of the total signal rate r the relevant parameter for the analysis is the overall signal-to-background ratio

$$\rho = \frac{\sum_{ki} s_{ki}}{\sum_{ki}(s_{ki} + b_{ki})}\,, \tag{7.6}$$

computed with hypothesis (B). The optimization criteria are the same as in section 2.2. Again the confidence levels are invariant if the weights are multiplied with a constant factor. In addition, an arbitrary constant may be added to the weights since the total number of events is fixed to n_{obs}. Criteria (i) and (ii) then lead, with the modified constraint (7.5) and a final fixing of the arbitrary constants, to the weight definition

$$w_{ki} = \mathcal{N} \frac{(1-\rho) \cdot s_{ki} - \rho \cdot b_{ki}}{(1-\rho) \cdot s_{ki} + (2-\rho) \cdot b_{ki}}\,. \tag{7.7}$$

If the local signal-to-background-ratio is equal to the global one, $s_{ki}/b_{ki} = \rho/(1-\rho)$, one obtains $w_{ki} = 0$. Events in bins with a larger signal-to-background-ratio are considered as more signal like with positive weight. If $\rho \ll 1$ but locally $s_{ki} \approx b_{ki}$, equation (7.7) approaches equation (2.6). If $\rho \approx 1$ but locally $s_{ki} \approx b_{ki}$, maximal background counting with a limiting weight $w_{ki} = -\mathcal{N}$ is reached.

If equation (7.7) is used instead of (7.2) for the example analyzed above, one gets $CL_{sb}^{(shape)} = 0.09$ for a signal plus background interpretation of the data set, to be compared with a median expectation $E[CL_{sb}^{(shape)}]_b = 0.04$ from background events. The result for the background hypothesis is $CL_b^{(shape)} = 0.67$ for the data set, close to the median background value $E[CL_b^{(shape)}]_b = 0.5$, while $E[CL_b^{(shape)}]_{sb} = 0.96$ is expected from background and signal events. These results do not support the existence of a signal. In this case the background was probably underestimated.

This test on $CL_b^{(shape)}$ and $CL_{sb}^{(shape)}$ has, of course, less discrimination power than the analysis based on CL_b and CL_{sb}, but it is an important cross check if an indication for a signal is found.

The systematic errors can be studied as described in section 5 with the exception that the approximation (5.2) is not applicable. The full errors matrices $\sigma_{j,ki}^{(s)}$ and $\sigma_{j,ki}^{(b)}$ have to be known.

8. Summary

The method of fractional event counting for extracting confidence levels is presented, based on a weighted sum over the observed events as the test statistic. A simple weight function (2.9), depending on a discriminating variable ξ, is derived. It contains one free parameter R, proportional to the signal rate, whose choice depends on the statistical question to be answered. Several criteria for its optimization are discussed. If a theoretical signal rate r is known, meaningful R values range from 0 to r. When upper limits have to be computed, R should be chosen so as to give the best upper bounds expected from the background hypothesis. With a value $R = r/2$, fractional event counting is very similar to the likelihood ratio method and the results are almost identical, as long as the signal-to-background-ratio is not larger than 1. In the presence of very low background bins, the weight function presented here has the advantage that it avoids singularities in the test statistic. Correlated systematic errors are included in the computation of confidence levels, using a fast numerical procedure.

The Bayesian and the frequentist treatments of systematic errors are compared. Both approaches agree, for a fixed weight function, if systematic errors introduce shifts of the distributions of the test statistic without modification of its shape, and if the systematic errors are invariant against translations of the distribution of test statistic. The signal and background distributions of the discriminating variable may be arbitrary continuous functions and the distribution of the systematic errors may also be arbitrary.

It is shown how systematic errors can be incorporated into the weight definition. The algorithm presented drops insignificant bins and improves, at the same time, the discovery potential for a signal.

An observed excess of the test statistic over the background level may either be caused by an excessive number of events or by different shapes of the distributions of the discriminating variable ξ for data and background. An analysis is presented which extracts confidence levels for a given signal-to-background-ratio from a comparison of the predicted shapes of the ξ distributions for background and a hypothetical signal with experiment. It relies on the polynomial distribution and fixes the total theoretical number of events to the observation. Singularities in the test statistic can be avoided. If there is an indication for a signal, this additional test is a valuable, supplementary consistency check.

The problem of low rate experiments is investigated. It is already known that upper limits from analyses with discriminating variables are rather ambiguous since they depend on the choice of the weight algorithm and its parameters. To avoid subjectivity, it is proposed to use simply the signal-to-background ratio as the weight. The reliability of the computation of systematic errors was checked. It turns out that, for a low rate experiment with a large systematic uncertainty of the background, the probability for an upward fluctuation is always underestimated with the Bayesian treatment of the background error.

Acknowledgments

This work would not have been possible without many stimulating discussions in the Higgs working group of the OPAL collaboration and also in the LEP wide working group on Higgs searches. The author thanks U. Jost for carrying out many tests during the initial phase of this work and is indebted to D. Plane and J. von Krogh, whose suggestions significantly improved the readability of the manuscript. The author is pleased to acknowledge the Bundesministerium für Forschung und Technologie for the support given to the OPAL project.

A. Optimal weights depending on systematic errors

In the following it is shown why the solution of equations (6.1) is unique and the sensitivity of the analysis is always improved. Let N be the total number of histogram bins. The normalization condition $X_s = \sum_{ki} w_{ki} s_{ki}$=const. defines a $(N-1)$-dimensional hyperplane in the space of weights w_{ki}. The N inequalities $w_{ki} \geq 0$ define an $(N-1)$ hyper-planar object with N corners within this hyperplane, a so called simplex. The simplest examples are a connection line for $N = 2$, a triangle for $N = 3$ and a tetrahedron for $N = 4$. At the corners, only one of the w_{ki} is positive. The surface of the simplex consists of N hyper-planar objects of dimension $(N-2)$, which are simplices again. The simplest examples are the end points of the connection line for $N = 2$, the sites of the triangle for $N = 3$ and the surface triangles of the tetrahedron for $N = 4$. These surface elements are characterized by one vanishing w_{ki}. Two of the $(N-2)$-dimensional surface elements have one $(N-3)$-dimensional simplex in common. There are $N \cdot (N-1)/2$ of these objects, on which two weights vanish. This decomposition can be repeated until one reaches the corners. All curvature components on these substructures vanish.

The condition $\sigma^2/(\sum_{ki} w_{ki} s_{ki})^2 = p$ defines an N-dimensional hyper-ellipsoid, whose size depends on the constant p. For sufficiently small values of p all points of the simplex $w_{ki} \geq 0$ lie outside the hyper-ellipsoid. Because both the error ellipsoid and the simplex are convex and all curvature components of the ellipsoid are non-zero, there exists exactly one value of p for which the simplex becomes a tangential object of the hyper-ellipsoid. The coordinates of the tangential point are the desired weights. The point computed with (2.9) lies in the interior of the $N-1$-dimensional simplex. In general, the error ellipsoid containing it has a larger value of p than the solution of equation (6.1).

B. Comments on the comparison of two arbitrary hypotheses

In physical models changes of the parameters often induce local changes of rates with different signs. A simple example is the comparison of two angular correlations. In the following it is summarized how the weighting has to be modified to discriminate between two arbitrary models (A) and (B). Let a_{ki} and b_{ki} be the local rates. The previous results are reproduced with $a_{ki} = b_{ki} + s_{ki}$. The weight optimization can be repeated with the normalization $\sum w_{ki} \cdot (a_{ki} - b_{ki})$ =const., with the result

$$w_{ki} = \frac{\mathcal{N} \cdot U \cdot (a_{ki} - b_{ki})}{U \cdot a_{ki} + (1-U) \cdot b_{ki}}$$

where a free parameter U replaces R. To guarantee a positive denominator, U should be constrained to $0 \leq U \leq 1$. The weights can now become negative, but they have a lower and an upper bound. The folding procedures to get the distributions of the test statistic are the same, but X lies now in the interval $-\infty$ to ∞ and the lower integration limit in the confidence level integrals has to be set to a sufficiently large negative number.

The weighting with systematic errors leads to the system of linear equations

$$\begin{aligned} w_{ki} \cdot (a_{ki} \cdot k_1 + b_{ki} \cdot k_2) + \sum_{lm} w_{lm} \cdot \sum_j \sigma^{(a)}_{j,lm} \sigma^{(a)}_{j,ki} \cdot k_1 + \\ + \sum_{lm} w_{lm} \cdot \sum_j \sigma^{(b)}_{j,lm} \sigma^{(b)}_{j,ki} \cdot k_2 = a_{ki} - b_{ki} \end{aligned}$$

The numerical factors k_1, k_2 are defined as before. The requirement of positive weights is meaningless. It was introduced to circumvent bad knowledge of the spectral shapes of systematic errors in regions where the difference between the models is small. Here, bins with $a_{ki} \approx b_{ki}$ are not significant, and they can be dropped with the requirement that w_{ki} must have the same sign as $a_{ki} - b_{ki}$.

For the statistical test, based on the shapes of the distributions, equation (7.7) has to be replaced by

$$w_{ki} = \mathcal{N} \cdot \frac{a_{ki}/\sum_{ki} a_{ki} - b_{ki}/\sum_{ki} b_{ki}}{a_{ki}/\sum_{ki} a_{ki} + b_{ki}/\sum_{ki} b_{ki}}$$

This formula is symmetric and has no singularities. An example for its application is the above mentioned angular distribution check.

References

[1] F. James, L. Lyons and Y. Perrin (editors), *Proceedings of the workshop on confidence limits*, CERN, GENEVA, 17th-18th January 2000, *CERN yellow report* **2000-005** (2000) http://user.web.cern.ch/user/Index/library.html.

[2] *Fermilab workshop on confidence limits*, 27th-28th March 2000, http://conferences.fnal.gov/cl2k/.

[3] *Proceedings of the workshop on advanced statistical techniques in particle physics*, Durham (2002) http://www.ippp.dur.ac.uk/Workshops/02/statistics.

[4] L. Lyons, R. Mount and R. Reitmeyer (editors), *Statistical problems in particle physics, astrophysica, and cosmology*, SLAC, Stanford 8th-11th September 2003 http://www-conf.slac.standford.edu/phystat2003.

[5] L. Lyons and M.K. Unel (editors), *Statistical problems in particle physics, astrophysics and cosmology*, Oxford, 12th-15th September 2005, World Scientific Publishing Co.

[6] A.L. Read, *Modified frequentist analysis of search results (the CL_s method)*, *CERN yellow report* **2000-005** (2000) 81;*Presentation of search results: the technique*, *J. Phys. G* **8** (2002) 2693.

[7] T. Junk, *Confidence level computation for combining searches with small statistics*, *Nucl. Instrum. Meth.* **A 434** (1999) 435.

[8] V.F. Obraztsov, *Confidence limits for processes with small statistics in several subchannels and with measurement errors*, *Nucl. Instrum. Meth.* **A 316** (1992) 388;
V. Innocente and L. Lista, *Evaluation of the upper limit to rare processes in the presence of background, and comparison between the Bayesian and classical approaches*, *Nucl. Instrum. Meth.* **A 340** (1994) 396.

[9] OPAL collaboration, K. Ackerstaff et al., *A search for neutral Higgs bosons in the MSSM and models with two scalar field doublets*, *Eur. Phys. J.* **C 5** (1998) 19 [`hep-ex/9803019`].

[10] ALEPH, DELPHI, L3 and OPAL Collaborations, *The LEP working group for Higgs boson searches*, *CERN-EP/98-046*, CERN, Geneva, 1998.

[11] ALEPH, DELPHI, L3 and OPAL Collaborations, *The LEP working group for Higgs boson searches*, *CERN-EP/99-060*, CERN, GENEVA, 1999.

[12] R.D. Cousins and V.L. Highland, *Incorporating systematic uncertainties into an upper limit*, *Nucl. Instr. Meth.* **A 320** (1992) 331.

[13] W.T. Eady, D. Drijard, F.E. James and B. Sadoulet, *Statistical methods in experimental physics*, North Holland Publ. Comp., Amsterdam (1971).

[14] OPAL collaboration, G. Abbiendi et al., *Search for the standard model Higgs boson with the opal detector at lep*, *Eur. Phys. J.* **C 26** (2003) 479 [`hep-ex/0209078`].

[15] ALEPH, DELPHI, L3 and OPAL Collaborations, The LEP working group for Higgs boson searches, *CERN-EP/2003-011* (2003).

[16] E. Gross and P. Yepes, *SM Higgs boson hunting at LEP*, *Int. J. Mod. Phys.* **A 8** (1993) 407.

[17] J.F. Grivaz and F. Le Diberder, *On the determination of a mass lower limit for the Higgs boson in the presence of candidate events*, *Nucl. Instrum. Meth.* **A 333** (1993) 320.

[18] PARTICLE DATA GROUP collaboration, R.M. Barnett et al., *Review of particle physics. particle data group*, *Phys. Rev.* **D 54** (1996) 1.

[19] O. Helene, *Upper limit of peak area*, *Nucl. Instrum. Meth.* **212** (1983) 319.

[20] G. Zech, *Upper limits in experiments with background or measurement errors*, *Nucl. Instrum. Meth.* **A 277** (1989) 608.

[21] H. Hu and J. Nielsen, *Analytic confidence level calculations using the likelihood ratio and Fourier transform*, *CERN yellow report* **2000-005** (2000) 109.

[22] G.J. Feldman and R.D. Cousins, *A unified approach to the classical statistical analysis of small signals*, *Phys. Rev.* **D 57** (1998) 3873 [physics/9711021].

[23] B.P. Roe and M.B. Woodroofe, *Setting confidence belts*, *Phys. Rev.* **D 63** (2000) 13009.

[24] G.Punzi, *Ordering algorithms and confidence intervals in the presence od Nuisance parameters*, in ref. [5].

[25] F. Tegenfeldt and J. Conrad, *On Bayesian treatment of systematic uncertainties in confidence interval calculation*, *Nucl. Instrum. Meth.* **A 539** (2005) 407.

[26] W.A. Rolke, A.M. Lopez and J.Conrad, *Limits and confidence intervals in the presence of nuisance parameters*, *Nucl. Instrum. Meth.* **A 551** (2005) 493.

[27] L. Lista, *Including Gaussian uncertainty on the background estimate for upper limit calculations using Poissonian sampling*, *Nucl. Instrum. Meth.* **A 517** (2004) 360.

[28] S.I. Bityukov, *Signal significance in the presence of systematic and statistical uncertainties*, *JHEP* **09** (2002) 060 [hep-ph/0207130].

[29] OPAL collaboration, G. Abbiendi et al., *Flavour independent search for Higgs bosons decaying into hadronic final states in* e^+e^- *collisions at lep*, *Phys. Lett.* **B 597** (2004) 11 [hep-ex/0312042].