DEUTSCHES ELEKTRONEN-SYNCHROTRON

DESY 95-113
June 1995

Comparing Statistical Data to Monte Carlo Simulation - Parameter Fitting and Unfolding

G. Zech
Fachbereich Physik, Universität Siegen

ISSN 0418-9833

NOTKESTRASSE 85 - 22607 HAMBURG

To be sure that your preprints are promptly included in the
HIGH ENERGY PHYSICS INDEX,
send them to (if possible by air mail):

DESY
Bibliothek
Notkestraße 85
22607 Hamburg
Germany

DESY-Zeuthen
Bibliothek
Platanenallee 6
15738 Zeuthen
Germany

DESY 95-113
June 1995

ISSN 0418-9833

Comparing statistical data to Monte Carlo simulation - parameter fitting and unfolding *

G. Zech
FB-Physik, Universität Siegen
D-57068 Siegen
E-Mail: Zech@HRZ.UNI-SIEGEN.DE

June 9, 1995

Contents

*Work supported by Bundesminister für Forschung und Technologie (FK 056Si79I)

1 Introduction

The statistical analysis of data is an important part of most experiments in nuclear and particle physics. Some decades ago physicists were usually well educated in basic statistics in contrast to their colleagues in social and medical sciences. Today the situation is almost reversed. Very sophisticated methods are used in these disciplines, whereas in particle physics standard analysis tools available in many program packages seem to make a knowledge of statistics obsolete. This leads to strange habits, like the determination of the r.m.s of a sample through a fit to a Gaussian. More severe are a widely spread ignorance about the (lack of) significance of χ^2 tests with a large number of bins and missing experience with unfolding methods.

There exist many good monographs on statistical methods in data analysis [1, 2, 3, 4, 5, 6]. It is not intended to compete with these, but to concentrate on an aspect which is rarely discussed, namely the fact that in modern experiments acceptance and resolution have to be corrected through a comparison of experimental data with Monte Carlo simulations.

The purpose of a measurement is usually to verify a theory, to determine one or several unknown parameters, or, if little or nothing is predicted, to measure a physical quantity or a distribution of it.

The first case - testing a hypothesis - is the simplest. It will be treated in chapter 3, where we discuss the usual χ^2 comparison of the measurement and the simulation. We also sketch empirical distribution function (EDF) tests like the Kolmogorov-Smirnov test, which are not as much appreciated by particle physicists as they should. In this chapter we also sketch the statistics of weighted events, a tool that is also needed in the following chapters.

In the forth chapter we present the standard method to fit Monte Carlo distributions to data using the least square and maximum likelihood methods. During the parameter iteration process the Monte Carlo distributions have to be modified. It is shown how this can be done by weighting the events thus avoiding to repeat the generation. Sometimes it is possible to use moments or other estimators to infer parameters from a sample. Examples for an efficient use of this method are given. Finally a technique to reduce the number dimensions in multivariate distributions without loss of information is discussed.

The fifth chapter deals with the more complex problem of unfolding. The standard least square unfolding method is closely related to parameter fitting, however in addition one has to deal with oscillations of the unfolded distributions. Several different unfolding techniques and regularization schemes are discussed, including iterative and binning free methods.

In chapter 6, finally, we study confidence intervals and discuss the computation of upper and lower limits from a Bayesian point of view.

Throughout this article the emphasis is put on applications. The reader is assumed to be familiar with basic statistics. We will study simple examples, mostly one-dimensional

distributions and one parameter fits to simplify the presentation. The generalization to the multivariate case and the determination of a set of parameters is straight forward and will be indicated where necessary.

This report will certainly contain errors, misleading statements, sections which are unclear and misprints. I would appreciate very much, if you could communicate them to me.

2 Some definitions and preliminaries

Some of the following definitions become relevant only in the subsequent chapters. For convenience we state them already here.

2.1 "True" and "observed" distributions

A density $f(x)$ of a variable x is measured by an apparatus with finite resolution. The probability density $f'(x')$ to measure the quantity x' is given by

$$f'(x') = \int_{\infty}^{\infty} t(x', x) f(x) dx \tag{1}$$

where we call t the transfer function which includes smearing and acceptance losses. (The convoluted variables and functions will always be marked with a prime.)

To infer the transfer function the detector response is simulated. Monte Carlo events are generated according to a "true" distribution $g(x)$, which is chosen close to the expectation for $f(x)$. The simulation of the detector including trigger and reconstruction produces events following an "observed" distribution $g'(x')$.

$$g'(x') = \int_{\infty}^{\infty} t(x', x)\, g(x) dx \tag{2}$$

The data analysis is based on a sample of N experimental events characterized by the values x'_i of the variable x' and a sample of M simulated Monte Carlo events characterized by pairs of variables x_i, x'_i. Thus the functions f' and g' are not known analytically, but only indirectly and with statistical uncertainties.

Normally it is cheaper to generate a Monte Carlo event than to collect a real event. Thus the number of simulated events will be higher than that of the experimental ones and the statistical uncertainty on their distributions will be correspondingly smaller. Ideally the Monte Carlo errors can be neglected. This simplifies the analysis considerably. Unfortunately, in most cases we will have to include the statistical uncertainties from the simulation .

Usually we combine events in bins (we will discuss exceptions later) of x or x', respectively. The content of bin μ is d_μ (m_μ) for the experimental (Monte Carlo) event sample, and d'_μ (m'_μ) for the corresponding measured histograms. The number B of x-bins may be different from the number B' of x'-bins. For the simulated events we know also the number $m'_{\mu\nu}$ of events generated in bin ν and observed in bin μ.

The integrals (1) and (2) become sums and the transfer function t and becomes a matrix $\mathbf{T}$, where $T_{\mu\nu}$ is the probability for an event in the true interval ν to be found in bin μ of the

smeared distribution.

$$T_{\mu\nu} = \frac{\int_\mu dx' \int_\nu dx\; t(x', x) f(x)}{\int_\nu dx\; f(x)} \tag{3}$$

$$\approx \frac{\int_\mu dx' \int_\nu dx\; t(x', x) g(x)}{\int_\nu dx\; g(x)} \tag{4}$$

(The integration limits are given by the bin boundaries. Throughout this paper we omit details of sums and integrals, where these are obvious from the context.)

The approximation (4) is the better, the smaller the bins and the closer the agreement of $g(x)$ and $f(x)$.

$$d'_\mu \approx \sum_\nu T_{\mu\nu} d_\nu \tag{5}$$

$$m'_\mu \approx \sum_\nu T_{\mu\nu} m_\nu \tag{6}$$

The Relations (5) and (6) suffer from statistical fluctuations, and (6) in addition from the approximation (4). The equalities are therefore only approximate.

An estimate $\hat{\mathbf{T}}$ of the transfer matrix $\mathbf{T}$ is obtained from the Monte Carlo simulation

$$\hat{T}_{\mu\nu} = m'_{\mu\nu}/m_\nu \tag{7}$$

In order to test whether the functions $g(x)$ and $f(x)$ agree it is not necessary to determine T explicitly. One has just to compare the "observed" distributions d'_μ and m'_μ.

If we know $f(x) = g(x, \lambda)$ up to an unknown parameter λ, we have to vary λ and together with it m'_μ until the agreement with d'_μ is optimum.

In the worst case $f(x)$ is completely unknown, then we have to unfold the observed distribution d'_μ. This can in principle be done by inverting the matrix T, but as will be shown below, this straight forward method, and also other unfolding recipes are not without problems.

2.2 Weighted events, equivalent number of events

Frequently we have to handle weighted events. In the old days of bubble chamber experiments, for instance, decay distributions were corrected by computing event weights from the potential flight length. Nowadays Monte Carlo simulations have replaced these weighting techniques, but there are still cases where weighting is useful, a common one is background subtraction using negative weights. Also Monte Carlo samples frequently consist of weighted events. Modifying weights helps to avoid the repeated generation of events.

The statistical error of a sum of N weighted events with weights w_i

$$n = \sum_{i=1}^{N} w_i \tag{8}$$

is

$$\delta n = \sqrt{\sum w_i^2} \tag{9}$$

We define a number $\tilde{n}$, the *equivalent number of events* which is the number of unweighted events having the same relative error as the weighted sum.

$$\frac{\delta\tilde{n}}{\tilde{n}} = \frac{1}{\sqrt{\tilde{n}}} = \frac{\delta n}{n} \tag{10}$$

We obtain

$$\tilde{n} = (\sum_i w_i)^2 / \sum_i w_i^2 \tag{11}$$

For example a mixture of 10 events with weight 1 and of 10 events with weight 2 has the same statistical significance as 18 (equivalent) unweighted events.

The concept of equivalent event numbers is discussed in more detail in the Appendix A. There we see that equivalent event numbers follow distributions which are very similar to the Poisson distribution. This property is very useful for the likelihood analysis of experiments with low event numbers.

3 Comparing a measured distribution to a simulated one

In this section we study *goodness of fit tests* without bothering whether a parameter has been adjusted or not. The purpose is less hypothesis testing but the detection of systematic errors. A comprehensive and rather complete review is given in Ref. [9] We start with an example and then consider the general problem.

3.1 A simple example

In Figure 1a we compare a measured histogram to a Monte Carlo prediction. For simplicity we assume that we can neglect the statistical fluctuations of the simulation. From a visual inspection of the plot we recognize a significant excess of Monte Carlo events at large x values and a corresponding deficit at the left hand peak.

3.1.1 The χ^2-test

Assuming that the simulation describes the data, the numbers d'_μ will follow Poisson distributions with mean m'_μ and variance m'_μ. (We neglect Monte Carlo fluctuations.) The χ^2 for the histogram (Fig. 1) is

$$\chi^2 = \sum_\mu \frac{(d'_\mu - m'_\mu)^2}{m'_\mu} \tag{12}$$

We get a value of 90 for 72 bins (NDF), which is perfectly acceptable, contrary to the visual impression. The corresponding χ^2-probability is 7 %.

What does this mean? By how much is the theoretical (Monte Carlo) distribution allowed to deviate from the data to be acceptable? In the Appendix B we estimate, that the minimum detectable systematic error α_0 is

$$\alpha_0 \propto \frac{B^{1/4}}{N^{1/2}} \tag{13}$$

where B is the number of bins and N the total number of events. A necessary condition for the validity of (13) is that the systematic deviation is not oscillating, but extends over many bins and that B is large enough to approximate the χ^2 distribution by a Gaussian.

From the Relation (13) we learn, that the significance of a χ^2 test decreases in most cases with increasing number of bins.

χ^2-tests with large number of bins have little significance. On the other hand strongly localized systematic deviations, - which rarely occur - are only detectable with not too wide bins.

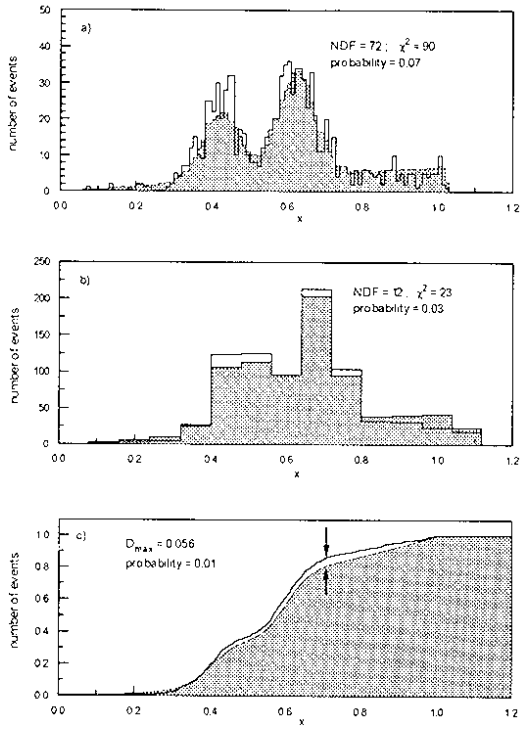

Figure 1: Comparison of experimental data to Monte Carlo simulation. In Figure a) and b) the chi squared is computed for 50 and 12 bin histograms, in c) the distributions functions are compared.

In our example the χ^2 test with 72 bins does not prove the deviation of the measured histogram from the data, which is obvious from the visual inspection. The reason for the failure of the χ^2 test lies in the fact that it does not take into account the same sign deviations of adjacent bins.

Figure 1b shows the same data sample as Figure 1a, but this time with the number of bins reduced to 12. The χ^2-value and the corresponding probability are now 23 and 3 %, indicating a more significant deviation of the measurement from the simulation.

The lesson from this example is not to use too large a number of bins for a goodness-of-fit test. Sometimes it is wise to try different binnings to search for different systematic problems. Usually physicists use the same binning to present the data, for parameter fits and for χ^2 tests. There is no need for this habitude. In our example one would prefer a relatively narrow binning for the presentation of the measurement but few bins for the test.

3.1.2 Kolmogorov-Smirnov-test

The problem of binning is avoided in a number of empirical distribution function (EDF) tests, the best known of these is the Kolmogorov-Smirnov-test.

The theoretical prediction is normalized to one and thus transformed into a probability density, the integral of which is the continuously rising distribution function $F(x)$.

$$F(x) = \int_{\infty}^{x} f(t)dt \tag{14}$$

The test statistic is the maximum difference between the experimental sum

$$S(x) = \frac{\text{number of events with } x_i < x}{\text{total number of events}} \tag{15}$$

and the prediction (14)

$$D_{max} = sup|S(x) - F(x)| \tag{16}$$

In our case $F(x)$ is approximated by the integrated and normalized Monte Carlo distribution $S_{MC}(x')$, defined analogously to (15). It is shown for our example in Figure 1c and compared to the experimental data $S(x')$.

The test quantity is the maximum deviation D_{max} of the two curves. For the integrated distributions we find $D_{max} = 0.057$ at $x_{max} = 0.71$. The probability to find $\sqrt{N}D_{max} > \alpha$ is for large N ($N \approx 100$) given by [1].

$$\lim_{N\to\infty} P(\sqrt{N}D_{max} > \alpha) = 2\sum_{m=1}^{\infty}(-1)^{m-1}e^{-2m^2\alpha^2} \tag{17}$$

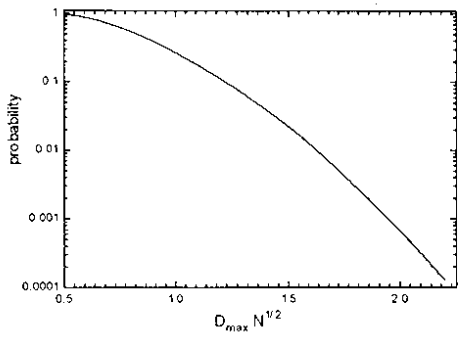

Figure 2: Conversion of Kolmogorov test statistic into probability.

Using this relation we can convert the observed deviation D_{max} into a probability. We obtain 2 % for the probability to observe in a similar experiment a larger deviation than the one actually found. This result demonstrates that the Kolmogorov-Smirnov test is more powerful than the χ^2 test in detecting deviations of the kind present in our example.

In general the Kolmogorov-Smirnov test is preferable to the more popular χ^2 test with its rather arbitrary binning. The asymptotic Relation (17) is a bit more complicated than the corresponding one for the χ^2 variable, but the probabilities are tabulated and thus easily available. We present the Relation (17) in Figure 2.

Unfortunately the application of the Smirnov-Kolmogorov test is restricted to univariate distributions. In case of multivariate distributions it can be applied to appropriate projections.

We will mention some other (more powerful) EDF tests in Section 3.3

3.2 The χ^2 test

When the statistical uncertainty of the simulation is not negligible we have to replace (12) by

$$\chi^2 = \sum_\mu \frac{(d'_\mu - m'_\mu)^2}{\delta^2_\mu} \tag{18}$$

where d'_μ and m'_μ are the observed experimental and Monte Carlo event numbers possibly including correction or weighting factors and δ^2_μ is the expected variance of the denominator assuming that the simulation and the experiment perfectly agree up to statistical fluctuations.

3.2.1 Statistical error of bin content

Below we give estimates of δ^2_μ. To simplify the notation we omit in this Section the bin index and the prime. We consider three different cases:

a) Negligible Monte Carlo error

In the trivial case, where the experimental events are not weighted and where the statistical error of the Monte Carlo simulation is negligible we have

$$\delta^2 = m \tag{19}$$

b) Monte Carlo errors included

The Monte Carlo prediction m is usually calculated by multiplying the number of events $\tilde{m}$ by a normalization factor c_N, usually smaller than one.

$$m = c_N \tilde{m} \tag{20}$$

We obtain from simple error propagation

$$\delta^2 = m + c_N m = m(1 + c_N) \tag{21}$$

c) Weighted events

Sometimes the experimental events and/or the simulation events are weighted and the bin content is a sum of event weights.

$$\begin{aligned} d &= \sum w_i \\ m &= \sum v_i \end{aligned}$$

Here the sums run over all events in the bin. Weighting is useful especially for Monte Carlo events. Repeated generation of events can be avoided.

As explained in Section 2.2 and Appendix A we define *equivalent event numbers* $\tilde{d}$ and $\tilde{m}$

$$\begin{aligned} \tilde{d} &= (\sum_i w_i)^2 / \sum_i w_i^2 \\ \tilde{m} &= (\sum_i v_i)^2 / \sum_i v_i^2 \end{aligned}$$

By error propagation we find for weighted events:

$$\delta^2 = \frac{md}{\tilde{d}} + \frac{m^2}{\tilde{m}} = m(\frac{d}{\tilde{d}} + \frac{m}{\tilde{m}}) \tag{22}$$

If the number of Monte Carlo events is smaller than the number of experimental events a better approximation is

$$\delta^2 = d(\frac{d}{\tilde{d}} + \frac{m}{\tilde{m}}) \tag{23}$$

3.2.2 Background subtraction

Often experimental data contain background which has to be estimated by a measurement (for instance beam gas reactions in storage ring experiments) or through a simulation. In both cases we obtain a certain number b of measured or simulated events which have to be multiplied by a normalization factor c_b and subtracted from the d_0 events observed in a bin

$$d = d_0 - c_b b \tag{24}$$

We treat the background events like normal events with negative weights of value $-c_b$ and compute the equivalent event number according to (11).

$$\tilde{d} = \frac{d^2}{d_0 + c_b^2 b} \tag{25}$$

Then we can use Equation (23) to compute the bin error δ.

3.2.3 Choice of bin width

It is not possible to establish generally valid rules for a optimum binning of data for a χ^2 test, however some advises can be given:

- The number of Monte Carlo events per bin should be large enough ($>\approx 20$) to approximate the Poisson distribution by a Gaussian and to replace the expectation in a bin by the observed number when estimating the error.
- Low event number bins should be combined with adjacent bins. In some textbooks it is proposed (originally by [11]) to choose bins of equal relative error. In the standard case with unweighted events this would mean that all bins contain the same number of Monte Carlo events. Tests indicate, that this procedure does not necessarily improve the power of the χ^2 test. (See also Section 3.4.)
- The χ^2 test relies on the assumption that the binning is chosen independently from observed deviations. Clearly, selecting a binning such that the χ^2 comes out low, strongly biases the result.
- Some reasonable choices for the number of bins are given in the Table 1:

Of course, these rules give only very rough guidance to a sensible binning in χ^2 tests and have to be used with common sense. One should also keep in mind that an optimum binning for adjusting parameters in a fit is usually not an optimum binning for a goodness-of-fit test. In the former case the experimental resolution and the structure, which one wants to resolve, play a decisive role, in the latter it is more the expected shape of an unknown background or a bias.

Table 1: Proposed numbers of bins for χ^2 tests

# of events	# of dimensions	# of bins
100	1	5
1000	1	10
10000	1	20
1000	2	25
10000	2	64
10000	3	125

In multivariate distributions with low event numbers it is often more informative to apply the χ^2 test to projections rather than to a multidimensional matrix.

In any case, test results should be interpreted with some scepticism. One should not consider a high χ^2 probability as a prove for the consistency of data with a prediction. A visual inspection of histograms often is more informative than a χ^2 test.

3.3 EDF-tests

3.3.1 Kolmogorov-Smirnov test and related supremum tests

The Kolmogorov-Smirnov-test has been conceived for the simple case with unweighted events and a theoretical prediction without uncertainty, which has been illustrated in the example we discussed at the beginning of this chapter. Of course, the concept can be extended to the more common problem, where we have to consider statistical uncertainties of the simulation and / or weighted events.

We can still form the normalized distribution functions of both, experimental and Monte Carlo events and measure their maximum distance D_{max}.

Since D_{max} scales with $1/\sqrt{N}$ in the asymptotic limit, and due to symmetry the number of Monte Carlo events M has to enter in exactly the same way as experimental number of events N, we find after some thinking that we have to replace $1/N$ by $1/N_{eff} = 1/N + 1/M$ when the events are unweighted.

For weighted events we help us again with the concept of equivalent event numbers and get the replacement

$$\frac{1}{N} \longrightarrow \frac{1}{N_{eff}} = \frac{1}{\tilde{N}} + \frac{1}{\tilde{M}} \tag{26}$$

and the probability

$$\lim_{N_{eff}\to\infty} P(\sqrt{N_{eff}}\, D_{max} > \alpha) = 2\sum_{m=1}^{\infty}(-1)^{m-1} e^{\ 2m^2\alpha^2} \tag{27}$$

There are several test quantities related to D_{max}. The maximum deviations $D_+ = max(S - S_{MC})$ and $D_- = max(S_{MC} - S)$ and the sum of these two $V = D_- + D_+$ are discussed in the literature. The first two D_- and D_+ are of minor importance but the latter V (introduced by Kuiper) should be used instead of D_{max} for distributions 'on a circle', where the zero is arbitrary as is the case for example for azimuthal distributions. It is also more powerful than D_{max} in detecting deviations in the width of resonances.

3.3.2 Tests based on quadratic statistics

The EDF-tests discussed so far measure essentially an excess or lack of events in a certain region of a distribution. In contrast to the χ^2 test they take correlations into account. However, the deviation is not really related to expected fluctuations. An excess (or lack) of events in the tails of a distribution has the same effect as an excess in a region where the density is high. This caveat is at least partially avoided in tests of the Cramer-von Mises family.

The test quantity Q is a quadratic function of the difference between the experimental and the theoretical distribution function.

$$Q = N\int_{-\infty}^{\infty}(S(x) - F(x))^2\,\psi(x)\,dF(x) \tag{28}$$

The tests of the quadratic class differ in the choice of $\psi(x)$. The Cramer-von Mises statistic, usually called W^2 uses $\psi = 1$.

$$W^2 = N\int_{-\infty}^{\infty}(S(x) - F(x))^2\ dF(x) \tag{29}$$

The choice $\psi = F/(1-F)$ corresponds to the Anderson-Darling statistic A^2.

$$A^2 = N\int_{-\infty}^{\infty}(S(x) - F(x))^2\,\frac{F(x)}{1-F(x)}\,dF(x) \tag{30}$$

In the literature it is stated, that two test statistics, W^2 and A^2 have similar power, but that A^2 is more sensitive to deviations in the tails of distributions [9]. We found that A^2 was the better choice in all examples investigated by us.

In Appendix A we summarize some useful relations and graphs for practical applications of EDF tests.

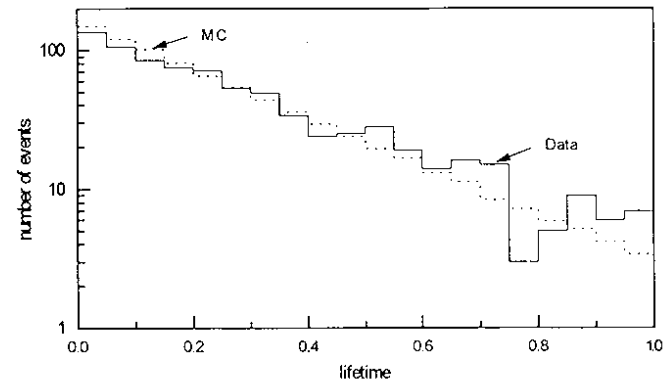

Figure 3: Exponential distribution (95 %) plus uniform distribution (5 %) compared to a purely exponential Monte Carlo simulation.

3.4 Comparison of tests

In Section 3.1.2 we have investigated a distribution with two peaks and a uniformly rising background which was more pronounced in the simulation. The results found there and from additional tests discussed in the previous section are summarized in Table 2.

To check the power of the various tests we have also compared an exponential distribution with 5 % uniform background to a purely exponential Monte Carlo distribution (see Figure 3). The results are given in Table 2.

The χ^2 test detects systematic errors only when the number of bins is small. The choice of bins with equal probability (e.p.) does not seem to improve the power of the test for small bin numbers. The statistic A^2 is in both examples the most sensitive test quantity.

3.5 Multivariate distributions

All the χ^2 related formulas we have presented above are also valid for multidimensional histograms, when the sums are running over all bins. Since the number of bins is usually larger in two or three dimensional histograms than in the case of a single variable, the χ^2 test is normally less significant. Also a visual illustration of a comparison of experimental and simulated data is more complicated.

Table 2: Test probabilities for Chi squared and EDF statistics for two examples.

distribution	test	probability
Fig. 1	χ^2, 72 bins	0.07
Fig. 1	χ^2, 12 bins	0.03
Fig. 1	D_{max}	0.018
Fig. 1	W^2	0.010
Fig. 1	A^2	0.006
Fig. 3	χ^2, 50 bins	0.10
Fig. 3	χ^2, 50 bins, e.p.	0.05
Fig. 3	χ^2, 20 bins	0.08
Fig. 3	χ^2, 20 bins, e.p.	0.07
Fig. 3	χ^2, 10 bins	0.06
Fig. 3	χ^2, 10 bins, e.p.	0.11
Fig. 3	χ^2, 5 bins	0.004
Fig. 3	χ^2, 5 bins, e.p.	0.01
Fig. 3	D_{max}	0.005
Fig. 3	W^2	0.001
Fig. 3	A^2	0.0005

In the two-dimensional case a useful technique is to plot or tabulate the χ^2 deviations as a function of the variables. To visualize correlated deviations of adjacent bins one should multiply χ^2 with the sign of the deviation $d_\mu - m_\mu$.

Also χ^2 tests of the projections of the distributions on one variable are quite informative.

It would be quite interesting to extend binning free tests from the simple one dimensional case to two or more dimensions. This proves to be quite complicated in practice, but may be possible in the future with the availability of powerful computers.

3.6 Summary

- The χ^2-test is feasible independent of the dimension of histograms. It can be applied without difficulty to adjusted histograms applying the usual *number of degrees of freedom = number of bins* minus *number of fitted parameters* rule. It is possible to handle weighted events and to take into account the simulation errors. The number of bins should be kept low.

- In the one-dimensional case EDF tests are to be preferred. They are free from the rather arbitrary choice of the bin widths and are more powerful in detecting systematic errors. The Anderson-Darling statistic A^2 is the optimum choice, but also the more intuitive Kolmogorov-Smirnov test gives good results.

- Background subtraction and event weights can be handled with the concept of equivalent event numbers.

- Test probabilities cannot replace a visual comparison of experimental and Monte Carlo histograms.

4 Inference of parameters

We assume that the function $f(x,\lambda)$ describing the data is known up to one unknown parameter λ. Our task is to determine λ and to estimate its error. Due to the measurement, $f(x)$ is transformed into an observed distribution $f'(x')$

We generate Monte Carlo events according to $f(x,\lambda_0)$ where λ_0 is a first guess for the true value of λ. The distorted Monte Carlo events, following $f'(x',\lambda_0)$, are compared to the measurement.

In principle, the best value of λ can be found by repeating the Monte Carlo simulation, modifying each time the parameter, until the agreement with the observed experimental data is optimum. This procedure is rather unpractical in situations where complicated detectors have to be simulated. It would require excessive computer time. (See section 4.4 for a detailed error discussion.)

An obvious possibility to avoid a comparison with the Monte Carlo simulation is to unfold the measurement and acceptance errors and to perform a fit of the theoretical distribution to the corrected data. This is bad practice, because the unfolding process has to be based on somewhat arbitrary assumptions, reduces the information and introduces correlations between adjacent points, which usually are neglected in the fit. Therefore we discard this method.

The alternative way is to modify the observed Monte Carlo distributions by re-weighting them with $f(x,\lambda)/f(x,\lambda_0)$, to compare it to the observed experimental distribution and to vary λ until the agreement is optimum.

In the next two subsection we discuss the *least square* and *maximum likelihood* fitting procedures with weighted events and in the following subsection we investigate the weighting procedures. In order not to complicate the discussion we neglect in Sections 4.1 and 4.2 additional fluctuations due to the re-weighting. Those will be treated separately in Subsection 4.4.

4.1 The least square method

We compare observed data and Monte Carlo histograms as described in Section 3.2.

Comparing the distributions we have to decide on their relative normalization, except in the very rare case, where we have an absolute prediction for the total number of events. There are two possibilities:

i) All Monte Carlo events are weighted by a common normalization factor $c_N(\lambda)$ such that their number agree with the total number of experimental events. The normalization factor c_N is recomputed each time the parameter λ is modified: $c_N = \sum d'_\mu / \sum m'_\mu$.

ii) The normalization factor is a free parameter in the fit.

Usually the results do not differ significantly in the two cases. The second possibility is somewhat simpler to implement, but is biased for small event numbers. (Due to the correlation of the width of the Poisson distribution with its mean value, high bin contents are preferred.) In the following we apply i) and assume that the number of simulated events in a bin follow a Poisson distribution, which in general is a very good approximation.

We have to minimize χ^2 with respect to λ.

$$\chi^2 = \sum_\mu \frac{\left(d'_\mu - c_N m'_\mu(\lambda)\right)^2}{\delta^2_\mu(\lambda)} \tag{31}$$

Here m'_μ is a sum of weighted events. The denominator is according to Relation (22)

$$\delta^2_\mu = \frac{c_N m'_\mu d'_\mu}{\tilde{d}'_\mu} + \frac{c_N m'^2_\mu}{\tilde{m}'_\mu} = c_N m'_\mu \left(\frac{d'_\mu}{\tilde{d}'_\mu} + \frac{m'_\mu}{\tilde{m}'_\mu}\right) \tag{32}$$

where $\tilde{d}$ and $\tilde{m}$ are equivalent event numbers. (See Section 3.2.1 and Appendix A).

When Monte Carlo errors can be neglected and the experimental events are unweighted the relation simplifies to

$$\delta^2_\mu = c_N m'_\mu \tag{33}$$

This approximation is justified if the number of Monte Carlo events is large compared to the experimental event number. If, for instance, there is a ratio of ten, then the Monte Carlo error is by a factor of $\sqrt{10}$ smaller than the error from the measurement. Since the errors enter in quadrature the contribution to the overall error squared is 10% and to the error about 5%. This numerical example tells us that one should try to have a ratio of ten or larger of the event numbers of Monte Carlo to data in order not to limit the precision of a measurement by the simulation.

As mentioned in Section 3.2.2, event weights are necessary when background has to be subtracted. The background can either be measured experimentally (beam gas reactions for example) or simulated. The background events are added with negative weights, which have to be properly normalized, to the data.

4.2 The maximum likelihood method

The likelihood estimation of parameters requires the knowledge of the probability density. We start with the simplest case, where Monte Carlo fluctuations are negligible.

The data d'_μ will be Poisson distributed in each bin with mean $c_N m'_\mu$ and variance $c_N m'_\mu$.

$$P(d'_\mu, c_N m'_\mu) = \frac{(c_N m'_\mu)^{d'_\mu}}{d'_\mu!} e^{-c_N m'_\mu} \tag{34}$$

The log-likelihood is

$$\ln L(\lambda) = \sum_\mu d'_\mu \ln(c_N m'_\mu) - c_N m'_\mu + const. \tag{35}$$

where the λ dependence is hidden in the prediction m'_μ and in the normalization c_N.

$$c_N(\lambda) = \sum_\mu d'_\mu / \sum_\mu m'_\mu \tag{36}$$

The parameter λ is found by minimizing the negative log likelihood.

Next we consider the case where the data consists of events of different weights. The distribution of a sum of weighted events is quite complicated. Fortunately it can quite well be approximated by a Poisson distribution using the concept of 'number of equivalent events' as defined above. This is shown in the Appendix A.

We obtain for the log likelihood

$$\ln L(\lambda) = \sum_\mu \tilde{d}'_\mu \ln(c_N \frac{\tilde{d}'_\mu}{d'_\mu} m'_\mu) - c_N \frac{\tilde{d}'_\mu}{d'_\mu} m'_\mu + const. \tag{37}$$

We have to complicate things further, to include also the statistical fluctuations of the Monte Carlo simulation. We have to introduce one new parameter per bin, the expectation value θ for the number of Monte Carlo events which gives - multiplied with the scale factor c_N - also the expectation for the number of equivalent Monte Carlo events. In the Appendix D we derive

$$\ln L = \sum_\mu \{\tilde{d}'_\mu \ln(c_N \frac{\tilde{d}'_\mu}{d'_\mu} \theta_\mu) - c_N \frac{\tilde{d}'_\mu}{d'_\mu} \theta_\mu + \tilde{m}_\mu \ln(\frac{\tilde{m}'_\mu}{m'_\mu} \theta_\mu) - \frac{\tilde{m}'_\mu}{m'_\mu} \theta_\mu\} \tag{38}$$

We now have to minimize -ln L not only with respect to λ but also with respect to the parameters θ_μ. The best values for θ_μ can be found analytically (see Appendix D):

$$\frac{\partial \ln L}{\partial \theta_\mu} = 0 \quad \longrightarrow \quad \theta_\mu = \frac{\tilde{d}'_\mu + \tilde{m}'_\mu}{c_N d'_\mu / d'_\mu + \tilde{m}'_\mu / m'_\mu} \tag{39}$$

This result can be inserted into (38) and the minimum as a function of the remaining parameter λ can then be computed numerically by a standard minimum searching program like MINUIT [12].

However for the computation of the errors also variations of the parameters θ_μ have to be included. This can be achieved by a two step procedure. In the first step the parameters λ are fitted from (38) with parameters θ_μ computed with (39). In the next step the parameters θ_μ will be left free in (38). If the starting values of the parameters are set to the values found in the previous fit only the errors will be computed.

4.3 Re-weighting the Monte Carlo events

So far we have not explained, how we obtain the Monte Carlo prediction $m'_\mu(\lambda)$.

4.3.1 Re-weighting individual events

We can modify the simulated distribution by re-weighting the events. Since we know for each event the true value x, we can obtain the "observed" distribution, which we compare to the data, $f'(x', \lambda)$ by weighting each event with $f(x, \lambda)/f(x, \lambda_0)$.

This procedure may still be rather slow, if many iteration steps are required in the fit. A faster method is illustrated by the following example.

4.3.2 Example: Slope of a linear function

We consider a simple example, where the function g is depending linearly on a parameter λ. Angular distributions often are of this form.

$$f(x; \lambda) = (1 + \lambda x)/(1 + \lambda/2) \quad \text{for } 0 < x < 1 \tag{40}$$

The Monte Carlo events are generated uniformly in x between 0 and 1, smeared according to the experimental resolution and histogrammed in x' bins yielding the distribution $m'_{0\mu}$. The same events are weighted by x and histogrammed into $m'_{1\mu}$. The two Monte Carlo distributions are shown in Figure 4a and b. The dotted histograms represent the distributions before smearing. In Figure 4c the experimental histogram d'_μ is shown.

It is compared to a superposition of the smeared Monte Carlo distributions:

$$m'_\mu(\lambda) = m'_{0_\mu} + \lambda m'_{1\mu} \tag{41}$$

Thus the smeared Monte Carlo distribution can be adjusted to an arbitrary value of λ without repeating the simulation, just by varying the relative weight of the two smeared distributions m'_0 and m'_1 and by adding them.

The dots in Figure 4c shows the proper superposition to fit an experimental histogram d'_μ. Also the corresponding superposition of the original "true" Monte Carlo distributions is given in the same figure.

4.3.3 Linear superposition of simulated distributions, Taylor expansion

Before we discuss the fitting procedure itself, we want to generalize our method. In the preceding example we were able to construct the Monte Carlo prediction by adding "ob-

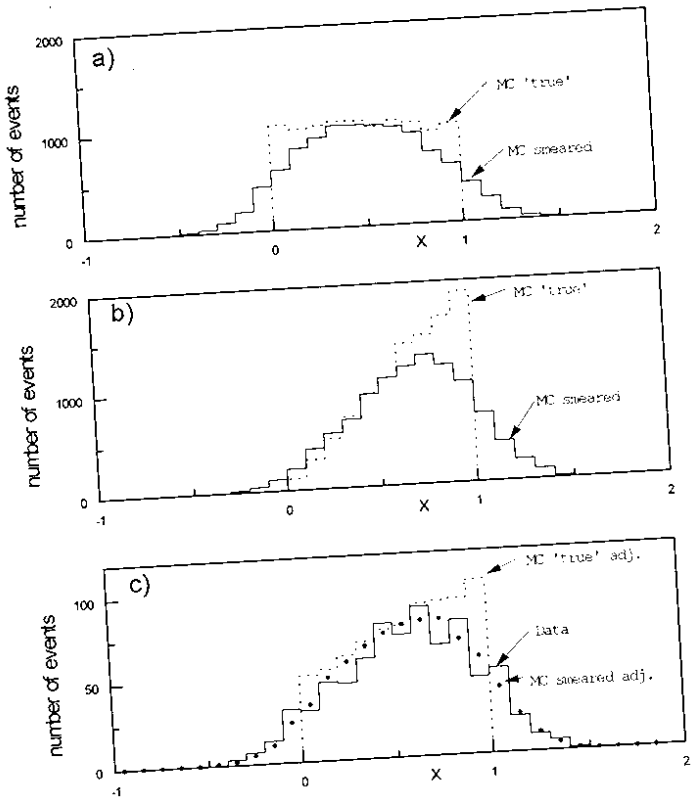

Figure 4: Fitting a superposition of smeared Monte Carlo distributions a) flat and b) triangular (weighted with x) to data c).

served" distributions. The relative weights of these distributions depended on the unknown parameter.

Obviously it is not essential that the function $f(x;\lambda)$ be linear in λ. It could be as well of the more general form

$$f(x;\lambda) = f_0(x)(1 + C_1(\lambda)f_1(x) + C_2(\lambda)f_2(x) + ...) \tag{42}$$

where C_i are arbitrary functions of the parameter λ. We would generate events according to $f_0(x)$ and fill the smeared events into the x' histogram m'_0. The same events weighted by the functions $f_1(x)$, $f_2(x)$, etc. are histogrammed into m'_1, m'_2, etc..Then the Monte Carlo prediction for bin μ is

$$m'_\mu(\lambda) = m'_{0\mu} + C_1(\lambda)m'_{1\mu} + C_2(\lambda)m'_{2\mu} + ... \tag{43}$$

which again can be computed for arbitrary values of λ from the smeared histograms. (Realize that the distributions m'_1, m'_2, .. corresponds to f_0 times f_1, etc.!)

In many applications the functions to be adjusted are not of the simple form (42), however, then we can perform a Taylor expansion of f in powers of $\Delta\lambda = \lambda - \lambda_0$ and bring it into the desired form.

$$f(x;\lambda) = f(\lambda_0) + \Delta\lambda\frac{df}{d\lambda}(\lambda_0) + \frac{\Delta\lambda^2}{2!}\frac{d^2f}{d\lambda^2}(\lambda_0) \tag{44}$$

$$= f(\lambda_0)\left(1 + \Delta\lambda\frac{1}{f}\frac{df}{d\lambda}(\lambda_0) + \frac{\Delta\lambda^2}{2!}\frac{1}{f}\frac{d^2f}{d\lambda^2}(\lambda_0) + ...\right) \tag{45}$$

On the right hand side of the equation the function and its derivatives have to be evaluated at λ_0. If λ_0 is a reasonably good estimate, one or two terms will be sufficient to approximate $f(x,\lambda)$.

4.3.4 Example: Lifetime fit

We illustrate this method with a lifetime fit [14]. We expand

$$g(t;\lambda) = \lambda e^{-\lambda t} \tag{46}$$

to second order and get

$$g(t;\lambda) = \lambda_0 e^{-\lambda_0 t}(1 + \frac{\Delta\lambda}{\lambda_0}(1 - \lambda_0 t) + \frac{\Delta\lambda^2}{\lambda_0^2}(-\lambda_0 t + \lambda_0^2 t^2/2) + ...) \tag{47}$$

Monte Carlo events are generated following the distribution $g_0(t) \propto \exp(-\lambda_0 t)$. A sample of events with "true" lifetimes t and measured values t' is obtained. The t' histogram of the events is denoted by m'_0. The histograms m'_1 and m'_2 are produced from the very same

events *weighted with their "true" lifetimes* t and the square t^2 respectively and *histogrammed in "observed"* t' bins.

In Figure 5b we show an experimental lifetime distribution measured with a resolution of $\tau/10$ and acceptance losses for small lifetimes. Figure 5a represents the Monte Carlo simulation $m'_0(\lambda_0, t')$ generated with a rough lifetime estimate of $\tau_0 = 1$ which is about 20 % shorter than the "true" lifetime (used to generate the data). Figure 5a also contains the distributions $m'_1(\lambda_0, t')$ and $m'_2(\lambda_0, t')$. The experimental distribution is than compared to the superposition

$$m'(t';\lambda) = a_N \left(m'_0(t')(1 + \frac{\Delta\lambda}{\lambda_0}) + m'_1(t')\lambda_0(\frac{\Delta\lambda}{\lambda_0} + \frac{\Delta\lambda^2}{\lambda_0^2}) + m'_2(t')\lambda_0^2\frac{\Delta\lambda^2}{2\lambda_0^2} \right)$$

where in the fit the normalization parameter a_N and the deviation $\Delta\lambda$ are adjusted. The result of a χ^2 fit is compared in Figure 5b to the data.

In our example the difference of the lifetimes in the Monte Carlo simulation and the measurement were rather large. With a better estimate or one iteration the quadratic term of the Taylor expansion could have been neglected.

4.4 The statistical error of the simulation

In Section 3.2.1 we have derived the relative error for the comparison of data and Monte Carlo event numbers in a bin. There is another contribution to the χ^2 variation that we have neglected so far. It is related to the variation of a parameter during the minimum search.

Since the correct treatment of statistical errors related to the Monte Carlo simulation in a fit is far from being trivial, one should always try to keep the Monte Carlo errors as small as possible.

4.4.1 The Poisson error

The observed number of simulated events m'_μ in a bin μ follow to a good approximation a Poisson distributions (for exceptions see below). If the events are weighted, we can approximate the distribution by an equivalent Poisson distribution as discussed above. The statistical error on a bin is easy calculable and also its contribution to the χ^2 value of a bin (see Section 4.1).

4.4.2 The error connected to parameter changes

In a χ^2 fitting procedure we usually assume that we can determine the one standard deviation error of a parameter from a χ^2 change of 'one' relative to the minimum. This change is small

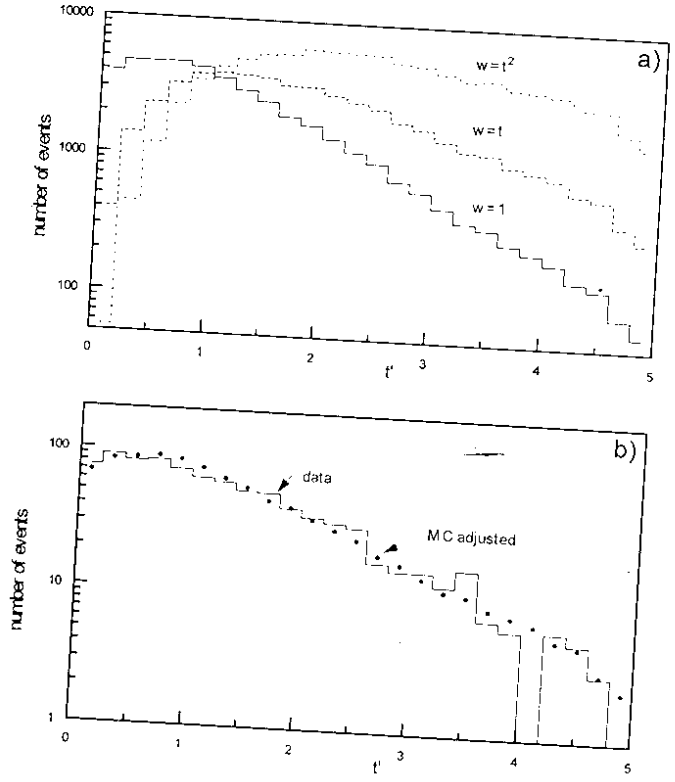

Figure 5: Fit of a lifetime distribution. The dotted histogram in b) is a superposition of the three histograms of a) with relative weights depending on $\Delta\lambda$ (see text).

compared to the statistical fluctuation of χ^2, which is $\sqrt{2NDF}$. This seems paradoxical at first sight, since two predictions which differ significantly, say by 3 st. dev. in a parameter may both have very acceptable χ^2 probabilities above 50 %. The obvious reason for the relatively good resolution of the parameter fit is related to the fact that most of the χ^2 fluctuation cancels when the *same* sample is compared to two predictions differing in the value of a parameter. The situation were completely changed if we used *different* data samples for the comparison.

The discussion above may seem purely academic, since in the fit always the same data sample is compared to the theory. However the Monte Carlo sample unavoidably is modified with each parameter change which is necessarily accompanied by a change of event weights.

We expect a purely statistical variation of χ^2 due to variations of the weights of

$$\delta\chi^2_w = r_{MC}\sqrt{2\sum_\mu(\sum_i \delta w_i^2/\sum_i w_i^2)} \tag{48}$$

Here r_{MC} is the contribution of the Monte Carlo fluctuations to the overall χ^2 at fixed λ, it is roughly the ratio of the number of experimental events to the sum of experimental and Monte Carlo events. The first sum runs over all bins (μ) and the second sum over all events (i) within that bin. The statistical fluctuation of the bin content is proportional to the sum of the quadratic changes $(\delta w_i)^2$ of the event weights w_i related to the parameter change (not to be confused with an uncertainty of the weights).

Let us illustrate the estimate (48) with a numerical example. We assume that we have 10 times more Monte Carlo events than experimental data, 50 bins and a relative change of the weights by 10 % when we change the value of the fit parameter. Then we find a small but non negligible value of $\delta\chi^2_w = 0.1$.

What can we do to reduce this unpleasant contribution? Unfortunately not much but increase the number of Monte Carlo events and reduce the number of bins. If the error $\delta\chi^2_w$ cannot be reduced sufficiently it has to be taken into account in the fitting procedure. This can be done analytically but a simpler way is to estimate its effect experimentally by subdivided the Monte Carlo sample into Z equivalent subsamples. The fit results from all subsamples are averaged and the mean quadratic deviation divided by Z is added quadratically to the parameter error given in an individual fit.

4.4.3 Fluctuations at the generator level

When we generate a certain number M of events according to a theoretical distribution and cast them into bins the bin contents follow in principle multinomial distributions. Usually the content of an individual bin is small compared to the total number of generated events. For this reason we can safely approximate the bin distributions by Poisson distributions.

The bin fluctuations in the "true" distributions can in principle be avoided. Since we know

the theoretical distribution we can throw away events such that the bin content corresponds exactly to the expectation. Alternatively one can use quasi random numbers for the event generation [16].

Of course this will also reduce the fluctuations in the distribution of the observed variables. The numbers m'_μ will follow multinomial distributions. The effect will depend on the acceptance and the resolution and will be sizable when the acceptance is large and the smearing is small compared to the bin size.

This variance reducing technique is of little practical importance. Physicists usually prefer to reduce the statistical error by increasing the number of Monte Carlo events and avoid the more complicated error handling for the multinomial distribution. However, if the statistical fluctuations of the simulation are neglected, then it makes sense to reduce the number of required Monte Carlo events by a controlled generation of the "true" distribution.

4.5 Sufficient estimators

In some cases a data sample can be represented by a single number (estimator, statistic), which contains the full information of the sample with respect to an unknown parameter of the corresponding distribution.

4.5.1 Example: Measurement of a lifetime

To illustrate this method we stick to the example, we just used to demonstrate the linear superposition method.

For a sample of N events with undistorted exponentially distributed lifetimes a *sufficient* estimator of the mean lifetime τ is the average $\bar{t}$ of the observed lifetimes. The statistic $\bar{t}$ contains the full information of the sample with respect to τ.

Clearly, for a distribution which is distorted by measurement errors and a finite acceptance $\bar{t'}$ will no longer be *sufficient* in the mathematical sense, but will still retain most of the information and be a sensible parameter to estimate the mean lifetime τ.

We determine $\bar{t'}$ and its error summed over all events

$$\bar{t'} = \sum t'_i/N$$

$$\delta^2\bar{t'} = (\overline{t'^2} - \bar{t'}^2)/N$$

with

$$\overline{t'^2} = \sum t'^2_i/N$$

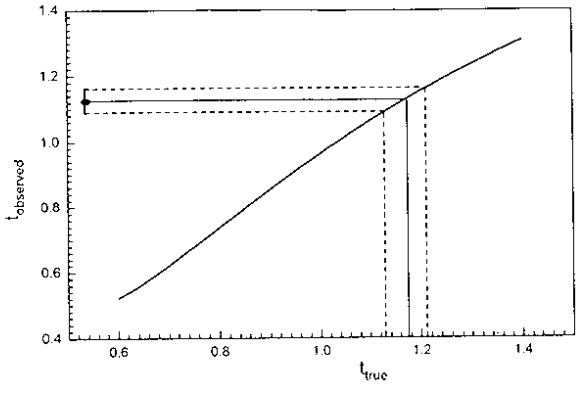

Figure 6: Transformation curve from observed mean lifetime to true lifetime computed by Monte Carlo simulation and application to a measurement.

From the Monte Carlo simulation we get the relation between $\overline{t'}$ and τ. This relation can be obtained by reweighting the Monte Carlo events as described above, instead of repeating the simulation many times. In Figure 6 we show the result of the simulation and the transformation of the measured average of t' to the parameter τ. The precision of the estimator method is almost identical to that of a least square fit.

4.5.2 General discussion

The determination of a parameter from an almost efficient estimator of a distribution has several advantages:

- No binning is needed.
- Problems related to low event numbers are avoided.
- The method is robust, easy to apply and very fast.

These properties make this method ideal for on-line applications, provided that a good estimator can be found.

In the example discussed above, we have used the first moment of a distribution to estimate a parameter. The moments method can be advantageous also in other cases, for example when the position and the width of a resonance are to be determined. There the first and the second moment are good estimators, whose biases can be removed using a correction curve computed from a Monte Carlo simulation.

Estimators are not necessarily moments of distributions. Ideally one should use the likelihood estimator, defined by the relations:

$$L(\lambda) = \prod f(x_i, \lambda) \tag{49}$$

$$\frac{d\,L(\lambda)}{d\,\lambda} = 0 \quad \rightarrow \quad \hat{\lambda} \tag{50}$$

The solution $\hat{\lambda}(x_i)$ of the second equation is a sufficient estimator. Using the last relation it is easy to prove that the mean of a sample following an exponential distribution is a sufficient estimator.

Unfortunately for most distributions an analytic solution cannot be found and numeric estimates are of little use. Sometimes an approximate expression for $\hat{\lambda}(x_i)$ can be obtained, as shown in the following example.

The error on the estimator can be determined, either by error propagation, or, 'experimentally' from several independent Monte Carlo simulations.

4.5.3 Example: Linear and quadratic distributions

Let us consider a sample of events distributed according to

$$f(x) = 0.5 + bx \;; \quad -1 \leq x \leq 1 \tag{51}$$

The log-likelihood is

$$\ln L = \sum \ln(0.5 + bx_i)$$

With b_0 being a rough estimate of b we set

$$b = b_0 + \beta \tag{52}$$

substitute it in the likelihood function, derive $\ln L$ with respect to β and set the result equal to zero to find the maximum.

$$\sum \frac{x_i}{0.5 + (b_0 + \beta)x_i} = 0$$

Neglecting quadratic and higher order terms in β and setting $f_{0i} = f(x_i, b_0)$ we find an estimate

$$\beta = \frac{\sum(x_i/f_{0i})}{\sum(x_i^2/f_{0i}^2)} \tag{53}$$

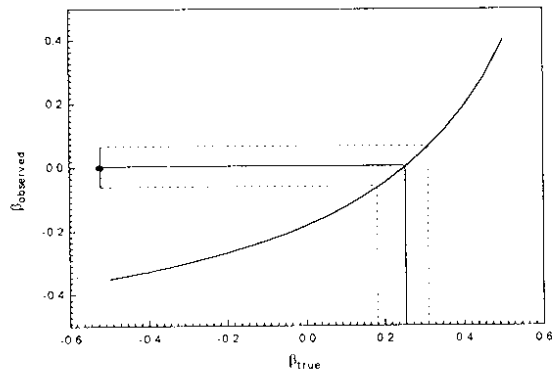

Figure 7: Transformation from almost sufficient estimator to slope of a linear distribution.

We have determined the slope of a linear distribution with a sample of 100 events generated with a parameter of $b_0 = 0.25$ using the estimator $\hat{\beta}$. The transformation curve from $\hat{\beta}$ to b is shown in Figure 7. Repeating the procedure 100 times and performing also a full likelihood fit for each sample we have checked that the resolutions of both methods agree within the statistical uncertainties. (The likelihood function is not parabolic, therefore the error given by standard fitting programs is much smaller than the r.m.s. error. We compared the r.m.s. errors.)

The resolution for the slope, computed from the first moment of a linear distribution is considerably worse, $\overline{x}$ is a very bad estimator in this case.

The linear estimator will still be almost sufficient when the distribution is distorted by measurement errors and acceptance, provided that the deformations are not excessive. Since likelihood fits for distorted distributions are not without problems, there the use of estimators has advantages.

For completeness we give also the estimators for the *quadratic distribution*

$$f(x) = a + bx + (1.5 - 3a)x^2 \qquad -1 \leq x \leq 1 \tag{54}$$

With

$$a = a_0 + \alpha \qquad b = b_0 + \beta \tag{55}$$

The estimators are

$$\hat{\alpha} = \frac{\sum(1 - 3x_i^2)/f_{0i}}{\sum(1 - 3x_i^2)/f_{0i}^2} \tag{56}$$

$$\hat{\beta} = \frac{\sum(x_i/f_{0i})}{\sum(x_i^2/f_{0i}^2)} \tag{57}$$

4.6 Reduction of variables

So far we only have considered single variable distributions. However all results can be applied equally well to multivariate distributions. The sum over all bins has just to cover all bins of the multidimensional histograms.

However multidimensional histograms often suffer from low event numbers per bin, which may lead to complications in the fitting. The χ^2 fit may be biased due to non-Gaussian errors and the maximum likelihood fit from the difficulties discussed in Section 4.3.

In some cases a substitution of variables accompanied by a reduction of their number can be applied without loss in information about the unknown parameters [15].

4.6.1 A simple example

Let us consider a simple probability density of two variables x_1 and x_2 with one unknown parameter λ

$$f(x_1, x_2; \lambda) = f_0(x_1, x_2)(1 + \lambda f_1(x_1, x_2)) \tag{58}$$

The substitution

$$u = f_1(x_1, x_2) \tag{59}$$

leads to a probability density for the single variable u.

$$h(u) = h_0(u)(1 + \lambda u) \tag{60}$$

Often the function $h(u)$ cannot be computed analytically, however this is not necessary if a Monte Carlo simulation is involved in the analysis. The generation of events can proceed using the full density (58). The observed variables x_1' and x_2' are inserted in (59) to compute u'. The histograms of u' for experimental and simulated events which retain the full information with respect to the unknown parameter are then used to perform the fit as described in the preceding subsections.

4.6.2 General case

We now turn to the general form of a probability density as treated in Section 4.5.2 but with several variables which we combine to a vector $\vec{x}$. The symbol λ may represent one parameter or a set of parameters.

$$f(\vec{x};\lambda) = f_0(\vec{x})(1 + C_1(\lambda)f_1(\vec{x}) + C_2(\lambda)f_2(\vec{x}) + ...) \tag{61}$$

If f is not a linear superposition of functions not depending on the parameters λ, it has to be expanded in a Taylor series as shown in Section 4.3.3 at an estimate λ_0.

With the substitutions

$$u_k = f_k(\vec{x}) \tag{62}$$

we transform the probability density into

$$h(\vec{u};\lambda) = h_0(\vec{u})(1 + C_1(\lambda)u_1 + C_2(\lambda)u_2 + ...) \tag{63}$$

The number of variables is reduced, if the number of variables originally is larger than the number of functions f_k. Again it is not necessary to perform the transformation analytically. It is done trivially by computing the new variables from the old ones for each event.

A good example, where this method is very useful is a measurement ratio λ of the vector to axial vector coupling constants in the reaction

$$e^+e^- \rightarrow \tau^+\tau^- \rightarrow e\mu + 4\nu \tag{64}$$

The decay density depends on the 6 momentum components of the final state leptons which can be replaced by 2 variables [15].

5 Unfolding

5.1 General remarks

In the previous Section we have shown how we can obtain a "true" distribution from an "observed" distribution through a comparison of the experimental data to a simulation. A necessary condition for the procedure was the possibility to parametrize the true distribution.

Can we compute the original distribution from the smeared one if we know nothing about its shape? The answer is "no". We have to make certain assumptions and for this reason unfolding is not a simple, straight forward procedure.

In Figure 8a we show a Gaussian superposed to a uniform distribution and in Figure 8b the corresponding smeared distribution. The Figure 8b contains also a second histogram (dotted) which is very similar to the first one. The original distributions (figure 8a), however, are quite different. It is obvious that unfolding the two smeared distributions will hardly reproduce the originals. If we replaced the single Gaussian by say ten well chosen Gaussians, the smeared distributions were indistinguishable within statistical uncertainties.The possibility to resolve narrow structures does not depend solely on the resolution but also on the available statistics. In principle with an infinite number of events arbitrarily small peaks can be resolved. A necessary condition is however that the resolution of the measurement is well known.

In Figure 9 we indicate graphically the effect of a mismatch of the experimental resolution (σ_{true}) and the resolution (σ_{MC}) used in the unfolding. We define

$$\delta = \frac{\sigma_{true} - \sigma_{MC}}{\sigma_{true}} \tag{65}$$

and

$$\sigma^2_{art} = \sigma^2_{true} - \sigma^2_{MC} \tag{66}$$

When we assume that the smearing is underestimated in the simulation a δ-line would have a width of σ^2_{true} after the measurement and an artificial width of σ^2_{art} after unfolding. The Table 3 shows that even a very modest mismatch of 10 % leads to an artificial width of almost 50 % of the resolution. Since the smearing function is hardly known to a better precision we have to be careful in the interpretation of unfolded distributions and should not expect miracles.

5.2 Empirical techniques

Before we discuss unfolding seriously we mention a widely spread empirical technique.

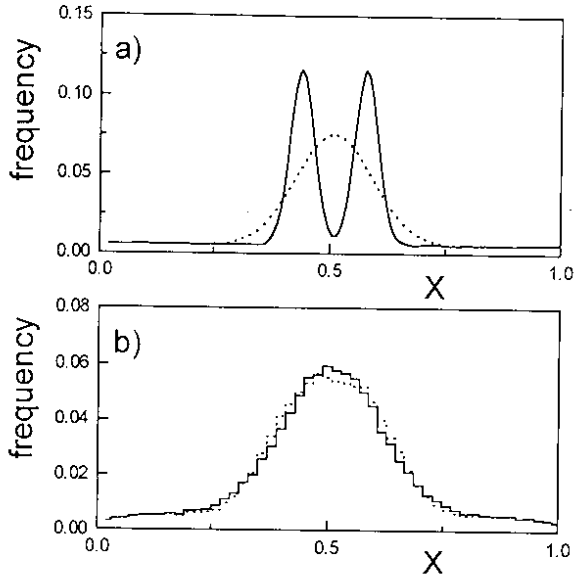

Figure 8: Effect of convolution for two different distributions. The original distributions (a) are smeared and compared two each other (b)

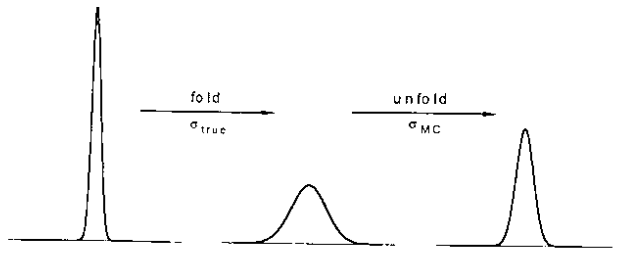

Figure 9: Effect of a 10 % mismatch between the resolutions used in smearing and unfolding of a Gaussian.

Table 3: Effect of mismatch between resolutions of smearing and unfolding.

δ	$\sigma_{art}/\sigma_{true}$
0.50	0.87
0.20	0.60
0.10	0.44
0.05	0.31
0.01	0.14

An estimate of the unfolded distribution is computed by multiplying each bin of the observed experimental distribution by the ratio of the "true" to the "observed" Monte Carlo event numbers (figure 10).

$$\hat{d}_\mu = \frac{d'_\mu}{m'_\mu} m_\mu \tag{67}$$

This method gives reasonable results, when i) the "true" experimental distribution is relatively well known and simulated, ii) biases of the observed variables with respect to the true variables are small and iii) the resolution is good compared to the bin width. The Monte Carlo simulation can be iterated, when the Monte Carlo distribution initially does not reproduce the data. The effect of biases can be reduced by choosing different bin boundaries in the observed and the "true" distributions, such that a maximum number of events from a true bin is found in the corresponding observed bin.

The simple technique is not appropriate for experiments, where peaks have to be resolved. In addition it has, at least in its primitive version, the caveat that it ignores events completely, when they fall outside the range of the "true" variables. Also the assignment of errors to the unfolded distribution is not without problems. The method may be used to get a first quick impression but is not appropriate for a scientific analysis.

5.3 Unfolding by matrix inversion

In Section 2 we have defined the transfer matrix T which relates the histograms of the true to those of the observed distributions. In vector notation the observed distribution $\mathbf{d}'$ is related to the "true" distribution $\mathbf{d}$ by

$$\begin{aligned} \mathbf{d}' &= T \cdot \mathbf{d} \\ T^+ \cdot \mathbf{d}' &= T^+ T \cdot \mathbf{d} \\ (T^+ T)^{-1} T^+ \cdot \mathbf{d}' &= \mathbf{d} \end{aligned}$$

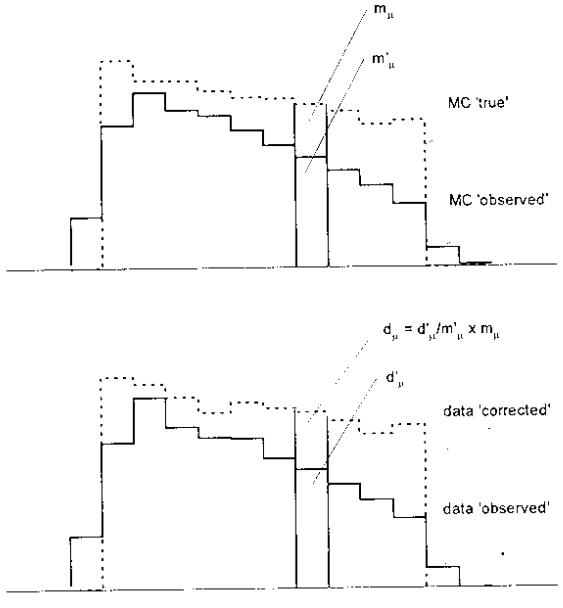

Figure 10: Primitive unfolding with ratio method

Using the Monte Carlo estimate $\hat{T}$ of T we get an estimate $\hat{\mathbf{d}}$ of $\mathbf{d}$.

$$\hat{\mathbf{d}} = (\hat{T}^+\hat{T})^{-1}\hat{T}^+ \cdot \mathbf{d}' \tag{68}$$

For a quadratic transfer matrix this simplifies to

$$\hat{\mathbf{d}} = \hat{T}^{-1} \cdot \mathbf{d}' \tag{69}$$

The uncertainties are obtained by simple error propagation. We set

$$\hat{S} = (\hat{T}^+\hat{T})^{-1}\hat{T}^+ \tag{70}$$

and get for the covariance matrix of d.

$$\begin{aligned} C_{ii} &= \sum_k \hat{S}_{ik}^2 d'_k \\ C_{ij} &= \sum_k \hat{S}_{ik}\hat{S}_{jk}^2 d'_k \end{aligned}$$

The matrix inversion method is a perfectly valid solution to the unfolding problem. It is, however, not much appreciated by most physicists, because the bin contents d_ν tend to vary strongly from bin to bin. The reason for this behavior is obvious from our discussion above. The smeared distributions of two input distributions cannot be distinguished if they agree on a large scale of x but differ by oscillations on a "microscopic" scale much smaller than the experimental resolution.

An example, taken from reference [18] is shown in Figure 11.

The covariance matrix shows a strong negative correlation between adjacent bins, indicating that the oscillations are statistically not significant. Nevertheless the deconvolution by matrix inversion may very well be used to compare data to a theoretical prediction, if the full error matrix is taken into account.

The matrix inversion method is rarely used, because a graphical representation of the result in many cases is difficult, if not impossible. Also the publication of large error matrices, is not very attractive. These problems can partially be reduced by choosing wide bins, but clearly this has to be paid for by a significant loss of information.

To avoid these problems, the oscillations have to be damped. This can be done by various regularization schemes which, however, are to a certain extend arbitrary and based on the prejudices of the analyzing physicist.

5.4 Least square and maximum likelihood methods

These are the standard methods. The contents of the bins of the unfolded distribution are considered as free parameters in a fit. The fitting proceeds along the lines presented in

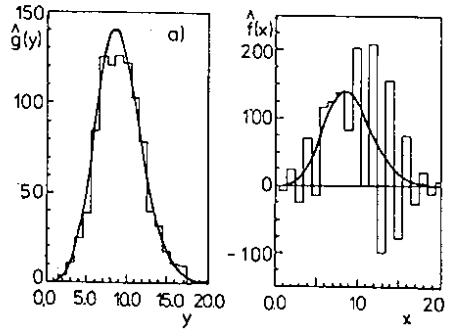

Figure 11: Unfolding by matrix inversion. (Example taken from Ref [18]). The result of the unfolding is shown in the right hand picture.

Section 4 and is especially simple, since the observed distribution is a linear superposition of the smeared distributions of "true" bins. To avoid oscillations, a regularization term is added to the purely statistical χ^2 or $-\ln L$ function.

Figure 12 is a graphical illustration of the fitting method.

The "observed" Monte Carlo distributions for events generated in "true" bins 1, 2 and 3 are presented in the second row. The experimental distribution in row 3 is fitted to a superposition of the three observed Monte Carlo distributions yielding the parameters λ_1, λ_2 and λ_3, the fractions of Monte Carlo events of the three classes which are needed to describe the data. The parameters λ_i can also be considered as weights that have to be applied to the "true" Monte Carlo distribution to get the "true" experimental distribution.

5.4.1 Least square fitting

We now formulate these ideas in a more mathematical form and define a χ^2 variable, containing a statistical term χ^2_{stat} and a regularization term R_{regu}.

The events $m'_{\mu\nu}$ generated in "true" bin ν and observed in bin μ are multiplied by λ_ν and compared to the observed data:

$$\chi^2_{stat} = \sum_{\mu=1}^{B'} \frac{(d'_\mu - \sum_\nu \lambda_\nu m'_{\mu\nu})^2}{\delta^2_\mu} \qquad (71)$$

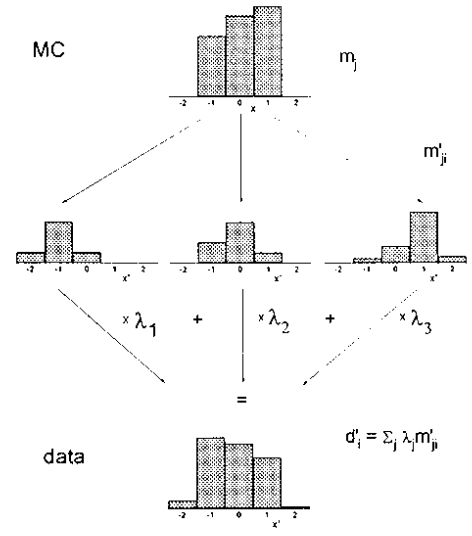

Figure 12: Illustration of fitting methods. Each "true" bin leads to a smeared Monte Carlo distributions. These are multiplied by factors λ which are adjusted such that the superposition matches the experimental data.

The error δ contains the statistical fluctuation of both the experimental data and the simulation and is computed as described in Section 4. We will list the results in the following paragraph.

To avoid statistically insignificant oscillations a regularization term

$$\chi^2 = \chi^2_{stat} + R_{regu} \tag{72}$$

is added and χ^2 is minimized. Thus estimates $\hat{\lambda}$ of the parameters λ are obtained which in turn are used to estimate the "true" experimental distribution.

$$\hat{d}_\nu = \hat{\lambda}_\nu m_\nu \tag{73}$$

5.4.2 Estimation of the covariance

We follow Section 4 to estimate the expectations δ^2_μ of the squared differences between data and Monte Carlo. Since we have already derived most of the results there, we can be rather short.

We use the abbreviation

$$M'_\mu = \sum_\nu \lambda_\nu m'_{\nu\mu} \tag{74}$$

for the predicted number of events in bin μ. Fluctuations in the number of events m_ν generated in a "true" bin have no effect on the unfolding result. The events of a "true" bin are distributed according to a multinomial distribution into "observed" bins. Thus the numbers $m_{\nu\mu}$ will follow multinomial distributions. These can be approximated by Poisson distributions, if the acceptance is low. If the events, data or Monte Carlo, are weighted, we use the concept of equivalent event numbers (see Appendix A). For instance, if background has to be subtracted, the background events are added to the data with negative weights.

a) negligible Monte Carlo error, unweighted events

$$\delta^2_\mu = M'_\mu \tag{75}$$

b) negligible Monte Carlo error, weighted data events

$$\delta^2_\mu = M'_\mu \frac{d'_\mu}{\bar{d}'_\mu} \tag{76}$$

c) Poisson Monte Carlo error, weighted data events

$$\delta^2_\mu = M'_\mu \left(\frac{d'_\mu}{\bar{d}'_\mu} + \frac{M'_\mu}{\tilde{M}'_\mu}\right) \tag{77}$$

with

$$\tilde{M}'_\mu = \frac{(\sum \lambda_\nu m'_{\nu\mu})^2}{\sum \lambda^2_\nu m'_{\nu\mu}} \tag{78}$$

d) Poisson Monte Carlo error, weighted Monte Carlo events, weighted data events

The expression for $\tilde{M}'$ has to be replaced by

$$\tilde{M}'_\mu = \frac{(\sum_\nu \lambda_\nu \sum_i w_{\nu\mu i})^2}{\sum_{\nu,i} \lambda^2_\nu w_{\nu\mu i}} \tag{79}$$

where $w_{\nu\mu i}$ is the weight of the i-th event in the "observed" bin μ originating from "true" bin ν.

e) Multinomial Monte Carlo error, weighted data events

Now the individual χ^2 terms are correlated, we have to use the weighting matrix V which is equal to the inverse of the covariance matrix C. The matrix C consists of contributions from the data $C^{(D)}$ and from the simulation $C^{(MC)}$. The probability for an event from "true" bin k to be found in the "observed" bin i is m'_{kj}/m_k

$$\chi^2_{stat} = \sum_{i,j} (d'_i - M'_i) V_{ij} (d'_j - M'_j) \tag{80}$$

$$V = C^{-1} \tag{81}$$

$$C = C^{(D)} + C^{(MC)} \tag{82}$$

$$C^{(D)}_{ij} = \delta_{ij} M'_i \frac{d'_i}{\bar{d}'_i} \tag{83}$$

$$C^{(MC)}_{ij} = \sum_k \lambda^2_k m'_{ki} (\delta_{ij} - \frac{m'_{kj}}{m_k}) \tag{84}$$

f) Multinomial Monte Carlo error, weighted Monte Carlo events, weighted data events

The last expression has to be replaced by

$$C^{(MC)} = \sum_k \lambda^2_k \frac{m'^2_{ki}}{\tilde{m}'_{ki}} (\delta_{ij} - \frac{m'_{kj}}{m_k}) \tag{85}$$

5.4.3 Maximum likelihood fitting

Following Section 4 we can also adjust the parameters λ in a maximum likelihood procedure. We minimize

$$-\ln L = -\ln L_{stat} + \frac{1}{2} R_{regu} \tag{86}$$

We have introduced the factor $1/2$ to obtain the same regularization strength as in the least square case. The overall log-likelihood can be written in most cases as a sum over contributions from the observed bins:

$$\ln L_{stat} = \sum_{\mu} \ln L_{\mu} \tag{87}$$

We obtain for the following expressions for $-\ln L_{stat}$ (see Section 4).

a) negligible Monte Carlo error, unweighted events

$$-\ln L_{\mu} = -d'_{\mu} \ln(\lambda_{\nu} m'_{\nu\mu}) + \lambda_{\nu} m'_{\nu\mu} + const. \tag{88}$$

$$= -d'_{\mu} \ln M'_{\mu} + M'_{\mu} + const. \tag{89}$$

where M' is defined by (74).

b) negligible Monte Carlo error, weighted data events

$$-\ln L_{\mu} = -\tilde{d}'_{\mu} \ln(M'_{\mu} \frac{\tilde{d}'_{\mu}}{d'_{\mu}}) + M'_{\mu} \frac{\tilde{d}'_{\mu}}{d'_{\mu}} + const. \tag{90}$$

c) Poisson Monte Carlo error, weighted data events

$$-\ln L_{\mu} = -\tilde{d}'_{\mu} \ln \theta_{\mu} + \theta_{\mu} - \tilde{M}'_{\mu} \ln(Q_{\mu}\theta_{\mu}) + Q_{\mu}\theta_{\mu} + const. \tag{91}$$

with

$$Q_{\mu} = \frac{\tilde{M}'_{\mu}/M'_{\mu}}{\tilde{d}'_{\mu}/d'_{\mu}} \tag{92}$$

Here the new parameters θ_{μ} are the expectation values $E(\tilde{d}'_{\mu})$ for the observed number of equivalent events. The parameters λ_{ν} searched for are hidden in M'_{μ} and Q_{μ}. We have now one parameter λ_{ν} for each "true" bin and one parameter θ_{μ} for each "observed" bin which have to be optimized. The values θ_{μ} that maximize L are found by setting $\partial \ln L/\partial\theta_{\mu} = 0$

$$\theta_{\mu} = \frac{\tilde{d}'_{\mu} + \tilde{M}'_{\mu}}{1 + Q_{\mu}} \tag{93}$$

with $\tilde{M}'$ given by (78)

d) Poisson Monte Carlo error, weighted Monte Carlo events, weighted data events

$\tilde{M}'$ is computed with Relation (79)

e) Multinomial Monte Carlo error, weighted data events

Now the bins are correlated. The probabilities $\epsilon_{\nu\mu}$ that a Monte Carlo event generated in bin ν falls into bin μ are introduced as new parameters. They are constrained by

$$\sum_{\mu} \epsilon_{\nu\mu} \leq 1 \tag{94}$$

The likelihood to observe $m'_{\nu\mu}$ events is

$$\ln L^{(MC)}_{\nu\mu} = m'_{\nu\mu} \ln \lambda_{\nu}\epsilon_{\nu\mu} + (m_{\nu} - m'_{\nu\mu}) \ln(1 - \lambda_{\nu}\epsilon_{\nu\mu}) + const. \tag{95}$$

For the overall likelihood we derive

$$\ln L = \sum_{\mu} \{\tilde{d}'_{\mu} \ln M'_{\mu} \frac{\tilde{d}'_{\mu}}{d'_{\mu}} - M'_{\mu} \frac{\tilde{d}'_{\mu}}{d'_{\mu}} + \sum_{\nu} \ln L^{(MC)}_{\nu\mu}\} \tag{96}$$

with

$$M'_{\mu} = \sum_{\nu} \lambda_{\nu}\epsilon_{\nu\mu} m_{\nu} \tag{97}$$

One has to be quite courageous to try fitting all the parameters λ_{ν} and $\epsilon_{\nu\mu}$. For example, for 20 true bins and 30 observed bins we get a total of 620 parameters. A least square fit is certainly preferable if the event numbers are large enough. With powerful computers and by subdividing the histograms a likelihood fit may also be possible.

f) Multinomial Monte Carlo error, weighted Monte Carlo events, weighted data events

The expression (97) has to be replaced by

$$M'_{\mu} = \sum_{\nu} \lambda_{\nu}\epsilon_{\nu\mu} \frac{m_{\nu}^2}{\tilde{m}_{\nu}} \tag{98}$$

5.4.4 Regularization

The most popular regularization scheme suppresses strong curvatures with a regularization term proportional to the second derivative of the unfolded distribution.

$$R_{regu} = \propto (\frac{\partial^2 f}{\partial x^2})^2 \tag{99}$$

This translates into the following expression for distributions with bins of equal size.

$$R_{regu} = \tau \sum_{\nu=2}^{B-1} (2\lambda_{\nu} m_{\nu} - \lambda_{\nu-1} m_{\nu-1} - \lambda_{\nu+1} m_{\nu+1})^2 \tag{100}$$

It vanishes for a linear distribution. The value of the constant τ determines the degree of oscillation damping. For very large τ the result of unfolding would be a linear distribution. Usually one would choose r such that the contribution of R_{regu} to χ^2 is small compared to one (of the order of 10 % or less) such that the results is influenced by a small fraction of a standard deviation. Figure 13 illustrates the effect of the regularization term. The unfolded distributions for different choices of τ is shown. The regularization is too strong on the histograms c) and d), histogram a) is not regularized.

In some cases the rough shape of the 'true' distribution is known. Then the regularization should be applied to the deviation of the result from the Monte Carlo input distribution.

The Equation (99) can be extended to multidimensional distributions. In principle, different regularization constants can be selected for different dimensions.

The regularization (100) will move the values $\hat{d}_\nu$ out of their minimum of χ^2_{stat}. The shift will be the larger, the larger the statistical error is. Thus points at the border of a distribution which often suffer from low statistics due to the decreasing acceptance may be moved considerably. In extreme cases the fitted position of such a bin may mainly be determined by the linear extrapolation from the adjacent bins and hardly by the measurement. To avoid this effect the expression (100) may be normalized to the expected statistical fluctuation.

$$R_{regu} = \tau \sum_{\nu=2}^{B-1} \frac{(2m_\nu - m_{\nu-1} - m_{\nu+1})^2}{\delta_\nu^2} \tag{101}$$

Here δ_ν^2 is the expected statistical fluctuation of the nominator. We leave it to the reader to estimate its value for the different statistical conditions.

Another possibility is to use bins of equal statistical significance. Then (100) has to be corrected for the varying distance of the bin centers.

Furthermore different regularization constants may be used in regions where strong variations of d_ν are expected and in parts where the distribution is known to be smooth.

5.4.5 Bias due to binning

In the unfolding methods mentioned so far the transfer matrix is estimated through (7)

$$\hat{T}_{\mu\nu} = m'_{\mu\nu}/m_\nu$$

The fraction of events generated in true bin ν and found in the observed bin μ will depend on the distribution of events in the true bin. This dependence will be negligible if either of the following two conditions is satisfied: i) the difference in slope

$$\Delta s = \frac{1}{f}\frac{df(x)}{dx} - \frac{1}{g}\frac{dg(x)}{dx}$$

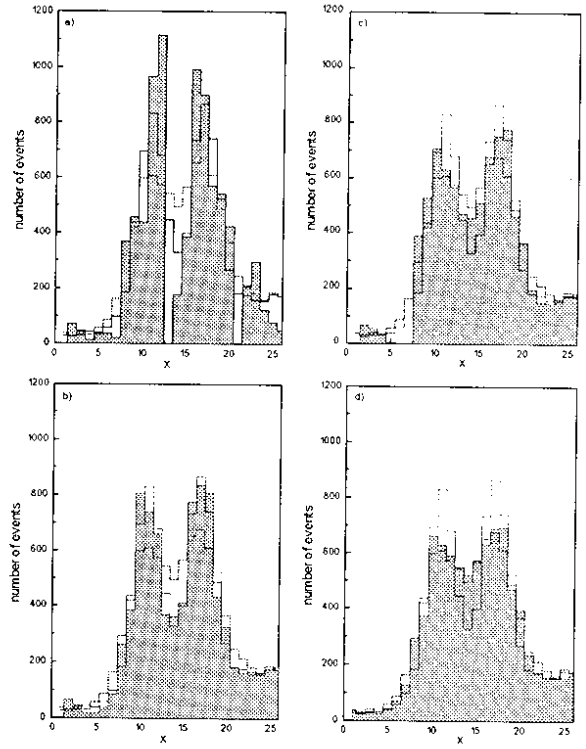

Figure 13: Unfolding with different regularization strengths. The regularization strength increases from a) to d). The solid (dotted) histogram represents the original (smeared) data, the shaded histogram is obtained by unfolding.

of the experimental and the simulated distributions is negligible, ii) the bin width is small compared to the experimental resolution, a condition which is not satisfied in most cases.

With some simplifications we can estimate the error introduced by a simulation function g different from the data distribution f.

We assume a Gaussian resolution σ, denote with Δx the difference between the bin centers x_ν and x'_μ of the true (ν) and observed bins (μ) and b the bin width. We find

$$\frac{\Delta T_{\mu\nu}}{T_{\mu\nu}} \approx \Delta s \left[\Delta x - \frac{\sigma^2}{b}(1 - e^{-\Delta x b/\sigma^2}) \right] \tag{102}$$

To get an idea for the order of magnitude of this error we set $\Delta x = \sigma$ and take the limit, where b is much smaller than the resolution σ. We get

$$\frac{\Delta T}{T} \approx \Delta s \; b/2 \tag{103}$$

The right hand side is the difference of the relative variations of the data and Monte Carlo functions within half a bin. For example a variation of f by 10 % within a bin and a constant g can introduce errors of the order of 5 %. A similar (slightly smaller) number is obtained when we choose $\Delta x = \sigma = b$.

To avoid this kind of error, one will try to simulate the data distribution as well as possible. If necessary the Monte Carlo distribution has to be iterated.

5.4.6 Other regularization schemes

The regularization method outlined above is not unique, in fact it is it is rather arbitrary. However the introduction of a personal prejudice into a statistical analysis is legal and common practice as long as the procedure used is well defined and documented. We will come back to this point in Section 6.

Generalizing the "curvature scheme" we can set the regularization term proportional to an arbitrary derivative squared

$$R_{regu} \propto (\frac{\partial^n f}{\partial x^n})^2 \tag{104}$$

or a combination of those. Setting $n = 1$ would favor a constant distribution and $n = 3$ a quadratic polynomial.

The second derivative is the most popular choice. Other regularization methods will be sketched below together with different unfolding schemes.

5.5 Some other unfolding methods

5.5.1 Blobel's method

This method [18] has several attractive features. The distribution is approximated by a superposition of cubic B-spline functions which is smooth by definition. Thus there are no steps at the bin boundaries, contrary to the standard least square method, described in Section 4.1 where we fit multipliers to the bin content. The B-splines are combined linearly to orthogonal functions with increasing frequency. This simplifies the minimization and offers a sensible way of regularization. Non-significant (high frequency) amplitudes are eliminated. Instead of cutting them off abruptly they are damped with a smooth function. The regularization is based on the 'second derivative suppression'.

Such a well defined procedure of course has also drawbacks. The decomposition of the function into orthogonal functions leads to a regularization with equal strength in the full variable range. Thus it is not possible to take variations of the statistical precision over a distribution into account. In most cases this is not a severe restriction.

5.5.2 Spectral window method

This scheme is described in refs.[18, 19]. It is closely related to the matrix inversion method but damps the oscillations in the same way as proposed in the method sketched above.

The inverse G^{-1} of the matrix

$$G = T^+ T \tag{105}$$

can be represented by the eigenvectors $g^{(l)}$ and eigenvalues γ_l of G.

$$G^{-1}_{nm} = \sum_l \frac{g^{(l)}_n g^{(l)}_m}{\gamma_l} \tag{106}$$

The unfolded histogram is obtained multiplying $T^+ \mathrm{d}'$ with G^{-1} (see Equation (68)). Obviously fluctuations in d' will be multiplied by γ^{-1}. Thus noise contributions essentially are due to small eigenvalues of G.

In the spectral window method the small eigenvector contributions are suppressed by a damping function:

$$\hat{G}^{-1}_{nm} = \sum_l f(\gamma_l) \frac{g^{(l)}_n g^{(l)}_m}{\gamma_l} \tag{107}$$

This function will be "one" for large components and smaller, or even zero for small components.

5.5.3 Cross entropy method

In this method [25] a 'entropy term' is added the standard least square expression. The entropy measures the fluctuations of the unfolded distribution integrated over the full variable space. For the typical unfolding problem this method is not advisable, because it ignores the importance of short distance fluctuations.

5.6 Iterative unfolding

5.6.1 Iterative method with binning

Matrix inversion can be done iteratively [23]. When this purely mathematical method is applied to the transfer matrix, of course the same fluctuating solution as discussed above is obtained. However, if the procedure is stopped after some iterations, fluctuations are suppressed. Thus regularization is inherent in this method. The smoothing is due to the slow transition from the starting distribution to the exact matrix inversion solution. The physicist's prejudice about the shape is hidden in the Monte Carlo input distribution, the degree of smoothing depends on the criterion used to stop the iteration process.

Before we write down formulas we will illustrate the method with a simple example, sketched in Figure 14. The histogram a) is the Monte Carlo input distribution, b) shows the observed distribution. The origin of the contributions to a certain bin are indicated. To get agreement with the data distribution c) the bins have to be scaled with the factors on top of the bins. These weights are then propagated back into the Monte Carlo "true" distribution. The result is the improved histogram d).

The corresponding equations are:

$$\lambda_\nu^{(k+1)} = \sum_\mu \hat{T}_{\nu\mu} \frac{d'_\mu}{m'^{(k)}_\mu} \tag{108}$$

with

$$m'^{(k)}_\mu = \sum_\nu \hat{T}_{\nu\mu} \lambda_\nu^{(k)} m_\nu \tag{109}$$

The convergence and the statistical significance is the better the closer it is to the experimental distribution. If no information is available it is recommended to start with an uniform Monte Carlo distribution.

In the limit $k \to \infty$ the weights $\lambda^{(k)}$ converge to the solution d_ν obtained by matrix inversion, provided that the latter does not contain unphysical negative parameters.

$$\lambda_\nu m_\nu = \sum_\mu \hat{T}^{-1}_{\nu\mu} d'_\mu \tag{110}$$

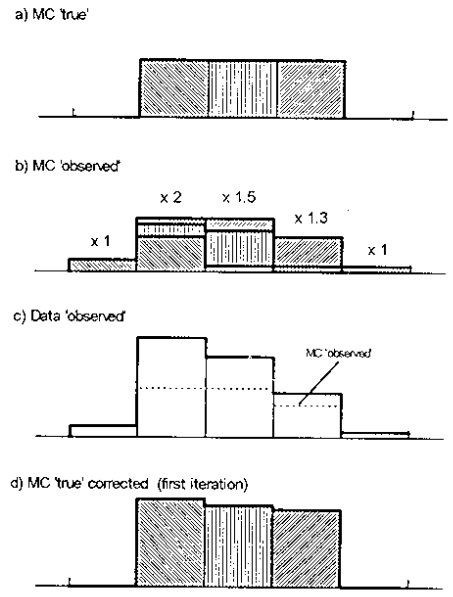

Figure 14: Illustration of iterative unfolding. All Monte Carlo events in an observed bin are weighted such that their sum matches the experimental data. The weights are propagated back into the original Monte Carlo "true" distribution.

This iterative method has been introduced in a different way in Refs. [20, 21, 22], where also its convergence is proven and was dicussed again recently in Ref. [?].

To find the optimum number of iterations, after each step the χ^2 is computed according to (72) and (99), the procedure is stopped after the step yielding the minimum.

This is illustrated with the following simple example where two Gaussians are superposed to a constant distribution. The original distribution for 5000 events is shown in Figure 15a, the smeared one in Figure 15b. A uniform Monte Carlo distribution of 10000 events was used to infer the transfer matrix.

The histogram obtained with the least square method without and with regularization is presented in Figures 15c, and 15d. The iterative result is given in Figure 15e.

The convergence of the iteration series can be accelerated by replacing the simple weight $d'_\mu/m'^{(k)}_\mu$ in (108) by the quantity $(d'_\mu/m'^{(k)}_\mu)^\alpha$ with $1 \leq \alpha < 2$. Large values of α can lead to oscillating weights, a number of 1.5 has given satisfactory results, reducing the number of steps by about 30 %.

A different iterative unfolding method is discussed in ref. [23].

In the iterative unfolding the statistical uncertainties of the simulation are not taken into account. Thus the number of Monte Carlo events should be much larger than the number of experimental events.

5.6.2 Iterative method without binning

Binning is always somewhat arbitrary and should be avoided where possible. Iterative unfolding can be generalized to a binning free method [26]. An individual weight is associated to each Monte Carlo event such that the weighted "true" distribution simulates the unfolded experimental data.

The unfolding without binning is free from the systematic biases, which are due to a non perfect simulation of the true function and which have been discussed in Section 5.4.5.

Another obvious advantages of the method is the possibility to choose the histogram parameters, selection criteria, and variables after the unfolding procedure. In this way the data analysis is much more flexible. If, for example, unfolding is done in two variables x and y, then also the unfolded r distribution ($r = -\sqrt{x^2 + y^2}$) is available.

A further advantage of binning free methods is the possibility to deal with low event numbers, which often cause problems when multivariate distributions have to be unfolded.

The extension of the standard iterative method to the unbinned case is straight forward. Instead of weighting all events in a bin each individual Monte Carlo event i at the position

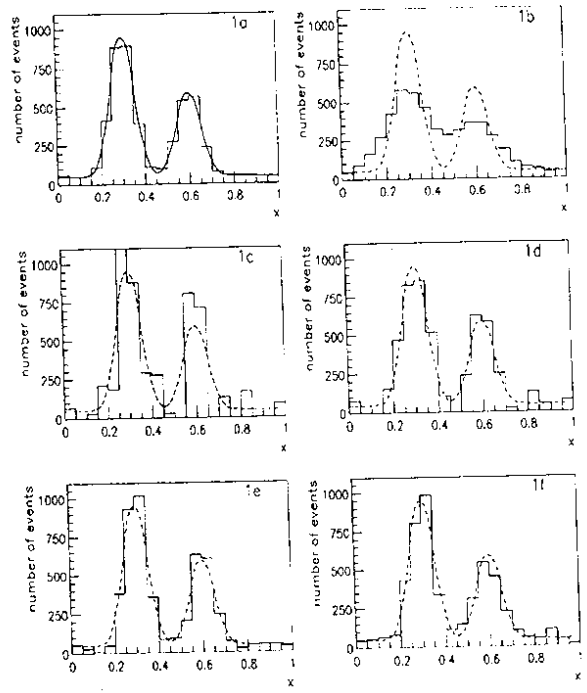

Figure 15: Comparison of unfolding methods.In a) and b) the original and the smeared distributions are shown. Least square unfolding with and without regularization is presented in c) and d), iterative unfolding with and without binning in e) and f).

x'_i is weighted according to the relative experimental and simulated local event densities $D'(x'_i)$ and $M'(x'_i)$.

$$w_i^{(1)} = \frac{D'(x'_i)}{M'(x'_i)} \tag{111}$$

The densities are estimated by counting the events in a region $x'_i - \Delta < x' < x'_i + \Delta$. The value of Δ is of the order of the resolution. It is not critical, because in the iteration procedure the weighted Monte Carlo and the experimental distributions become very similar. Alternatively the densities can be estimated from the x' range covered by the nearest N' events, where N' depends on the total event number and the resolution.

The x distribution of the weighted Monte Carlo events is the first iteration towards the unfolded distribution. The starting distribution for the next step should ideally consist again of unweighted events or events with weights depending only on x and not on x'. Thus we average the Monte Carlo weights over a certain x range.

$$\lambda_i^{(1)} = \frac{1}{N} \sum_{n=1}^{N} w_k^{(1)} \tag{112}$$

The sum extends over the N nearest neighbors of the event i in x space. (Alternatively a fixed region around x_i could be selected.)

The number of N (or equivalently the size of the x_i region) for the computation of the weight average is an important parameters. It steers the degree of smoothing in the unfolding. Its choice will depend on the available Monte Carlo event statistics and on the expected curvature of the true distribution.

The further iteration steps are obvious:

$$w_i^{(k+1)} = \frac{D'(x'_i)}{M'(x'_i)\lambda_i^{(k)}} \tag{113}$$

$$\lambda_i^{(k)} = \frac{1}{N} \sum_{n=1}^{N} w_k^{(k)} \tag{114}$$

The iteration is terminated by a χ^2 criterion identical to that described above. To perform the χ^2 comparison the events are grouped in bins. For the example described above the χ^2 is plotted as a function of the iteration number in Figure 16.

This procedure was applied to our example. For the computation of the densities and the weight averaging hundred adjacent events have been used.

The unfolded distribution, histogrammed in the same bins as above is shown in Figure 15f. The result is similar to those obtained with the conventional method. This is also seen in Figure ??, where the parameters of the two Gaussians derived from the three methods are compared. The procedures were repeated fifty times in each case to avoid accidental results.

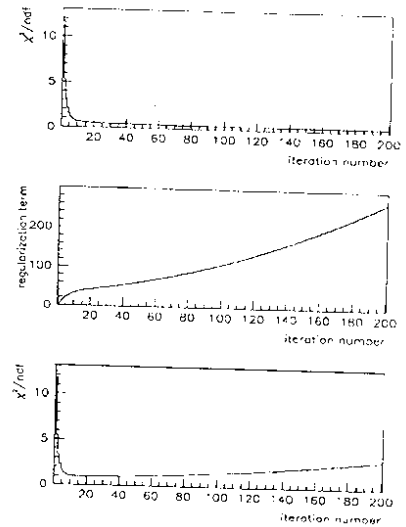

Figure 16: Chi squared as a funtion of the iteration number. The first plot shows the statistical term only, the second one the regularization term. The sum of both is given in the third plot.

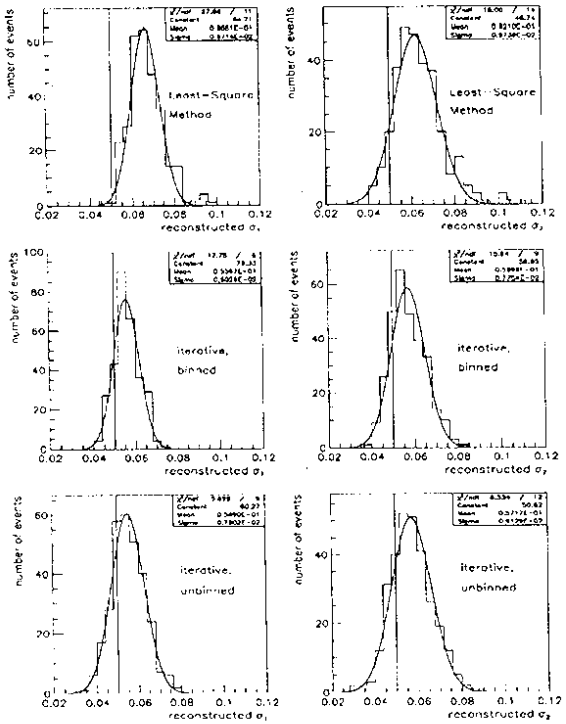

Figure 17: Comparison of the results of different unfolding methods. The widths of the two peaks of the previous figure as obtained in different unfolding procedures are shown. The true values are indicated by a line. The average unfolding result is larger due to regularization.

In our binning free method we compute the position uncertainty δ_{x_i} individually for each Monte Carlo event i from a sample of N neighboring events in x'

$$\delta^2_{x_i} = \sum_{n=1}^{N} (x_i - x'_i - x_n + x'_n)^2 / N \tag{115}$$

This quantity is independent of the experimental data but is depending to a certain extend on the shape of the Monte Carlo input distribution and in that sense on the prejudice of the physicist. (This dependence is rather weak.) The value δ_x is a measure of the correlations and is related to the range in x within which oscillations occur, if not suppressed by the regularization.

When we want to transform the unfolded event sample into a histogram we can compute the error matrix following the matrix inversion method (71).

5.7 Uncertainties related to the unfolded distribution

The errors of the unfolded distribution, including correlations, are given automatically in the matrix inversion and fitting methods. In other methods they can be deduced more or less indirectly. However, there is an unavoidable problem related to the regularization. The size of the errors depends on the degree of regularization. This is unfortunate, we would like to present uncertainties depending solely on the quality of the experimental data and not on degree of manipulation they have undergone.

As mentioned above, adjacent bins have negative error correlations: The χ^2 value does not change much when a bin is moved up and the two neighboring bins down. This is illustrated by the error ellipse in Figure 18. The regularization on the other hand tries to keep variations between adjacent bins small, leading to an χ^2 valley in the orthogonal direction along the diagonal. Even when the regularization is soft the combination of both reduces the area of the error ellipse considerably. This does not mean, however, that our knowledge has improved.

There is an obvious problem of representation of an unfolding result. Off-diagonal errors cannot be shown in a simple way in a graphical representation of an unfolded function. The diagonal elements of the error matrix depend strongly on the regularization.

Figure 19 illustrates the problem using a measurement of the structure function of the proton as an example. The diagonal errors as given from the χ^2 minimization are shown. The upper graphs are obtained without regularization. The errors and the fluctuations are large. Increasing the regularization constant transforms the same experimental data into a continuously rising function with much smaller error bars. Whereas in the upper plots the uncertainties are exaggerated they are underestimated in the lower ones.

Since the diagonal errors tell us little about the precision of the unfolded data, we have to find a way to present separately the two kinds of uncertainties associated to the unfolded

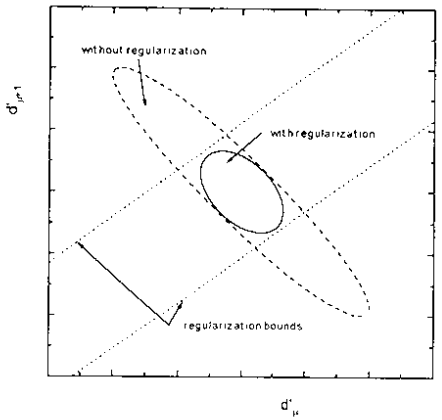

Figure 18: Correlation between two adjacent bins of the unfolded distribution. The area contained in the purely statistical error ellipse is reduced by the regularization bounds which disfavor large differences between adjacent bin contents.

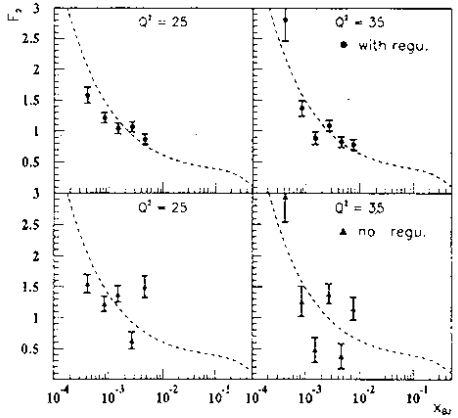

Figure 19: Unfolding of the proton structure function with and without regularization.

distribution, i) the statistical error related to the fluctuation of the number of events and ii) the error related to the smearing due to the finite resolution of the experiment, which leads to a negative correlation between neighboring bins. The degree of regularization is fixed by the analyzing physicist. The presented errors should not be influenced by this subjective element.

In order to present the errors graphically in such a way that both kinds of errors become transparent, we propose the following procedure:

i) The vertical error bar is computed from the statistical fluctuation of the number of events in the bin neglecting the variation due to bin to bin migration. This means that correlations between bins are ignored. These errors are similar to those of the simple ratio methods sketched in Section 5.2.

ii) A horizontal bar indicates the experimental resolution. The resolution can be estimated from the correlation obtained in a least square fit with no regularization or directly from the Monte Carlo simulation.

This presentation would indicate to the reader both, the statistical precision of the measurement, and its sensitivity to fast varying distributions.

6 Confidence limits, likelihood limits, upper and lower bounds

This section is only marginally related with the problem of comparing data with Monte Carlo simulation.

In fact upper limits for the expectation of Poisson distributed observations do not require simulations except for the normalization which can be corrected with standard methods and little computing power.

In contrast, the computation of confidence limits of parameters which are determined from a continuous distribution is not trivial. There a huge amount of simulation may be necessary (see Section 6.4). For this reason and since the concept of confidence limits is confronted with additional problems of more fundamental nature, we plea to replace it by the Bayesian concept of likelihood limits.

We will first repeat the definitions of confidence and likelihood limits, then we discuss some frequently occurring problems related to unphysical bounds and to background subtraction. We sketch the Monte Carlo corrections and finally draw conclusions.

6.1 Definition

When we measure or estimate a quantity x and find $\hat{x}_0$ we would like to give some kind of error limits which have a reasonable probability to contain the true value x_{true}. Classical statistics refuses to accept such probabilities but associates "objective" confidence limits x_{min}, x_{max} to the measurement.

Confidence intervals are defined as illustrated in Figure 20. Be α_1 (α_2) the probability to obtain a measurement $\hat{x}$ larger (smaller) or equal than $\hat{x}_0$ for the true value located at x_{max} (x_{min}). The quantity $c_l = 1 - \alpha_1 - \alpha_2$ is the confidence level.

The relation between the confidence limits, the distribution $f(\hat{x}, x_{true})$ and $\hat{x}_0$ is also shown in Figure 21. For given values of x_{true} the quantity x_{low} is computed such that the probability to observe $\hat{x} \leq x_{low}$ is α_2. In an analogue way x_{high} is defined through the probability α_1 to obtain $\hat{x} \geq x_{high}$. The two curves $x_{low}(x_{true})$ and $x_{high}(x_{true})$ contain the fraction $1 - \alpha$ ($\alpha = \alpha_1 + \alpha_2$) of the probability density $f(\hat{x}; x_{true})$. They can be computed analytically or by a Monte Carlo simulation and are independent of the measurement $\hat{x}_0$. The actual confidence limits for $\hat{x}_0$ are given by the intersection of the line $\hat{x} = \hat{x}_0$ with the two curves.

The definition given above is valid for continuous and discrete distributions of the observations $\hat{x}$. Confidence limits are invariant against variable transformations and thus an objective representation of the measurement. Note that it is not required that there exists any probability density for $\hat{x}$.

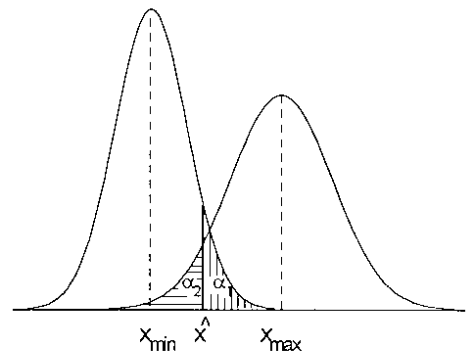

Figure 20: Illustration of the concept of confidence level. Note that the shape of the distribution may change when moving from the true value to x_{min} or x_{max}

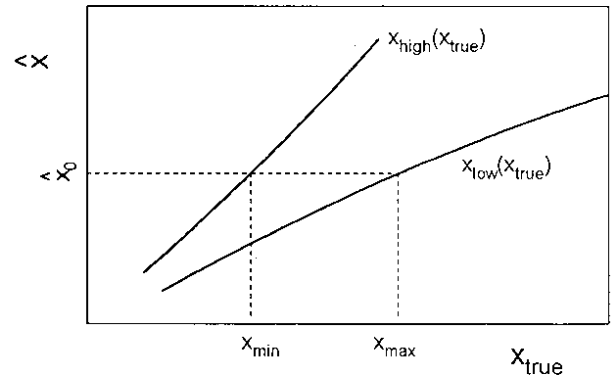

Figure 21: Confidence bounds. The bounds x_{low} and x_{high} confine a $1 - \alpha$ probability region for $\hat{x}$ inside as a function of the true value x_{true}. The confidence bounds x_{min} and x_{max} for the actual observation $\hat{x}_0$ are found from the intersection of $\hat{x} = \hat{x}_0$

When the shape of the distribution is independent of the location of the true value, the above definition simplifies to: α_1 (α_2) is the probability to obtain a measurement x smaller (larger) or equal than x_{min} (x_{max}) for the true value located at $\hat{x}_0$.

In principle one would like to choose the interval such that it has minimum length. In practice mostly symmetric intervals with respect to $\hat{x}_0$ are chosen. Another possibility is to select α_1 and α_2 equal. From these three choices only the last one is invariant against variable transformations.

Many measurements in particle and nuclear physics are mean values from a sample of observations. The one standard deviation error on the mean is computed from the root mean squared deviations. This is a very reasonable practice. When using the measured quantity to compute other variables depending on it, it is the r.m.s. error which is relevant in error propagation. It should be stressed that the r.m.s. error limits coincide with the 68.27 % confidence limits only when the observations follow a Gaussian with variance independent of the location of its central value.

Often one-sided confidence intervals (upper and lower limits) are given. An often occurring example is the 90 % upper limit $N < 2.30$ for a measurement of $\hat{N} = 0$ for a Poisson distributed number.

Confidence intervals in two or more dimensions can be defined in a similar way. For a measurement $(\hat{x}_0; \hat{y}_0)$ we can find a contour, such that for all true values lying outside, the probability to observe a value inside is less than α. However, when we are interested in only one variable, say x, we have to integrate over y, and we obviously need to know the probability density of y. Thus a purely classical solution to this problem does not exist.

6.2 Bayesian approach

The more modern Bayesian approach is more pragmatic and considerably simpler. Through Bayes theorem a probability density is associated to the unknown true value which is proportional to the likelihood function. (A uniform prior distribution is implicitly assumed for the parameter.)

$$f(\lambda) = \frac{L(x_1....x_N, \lambda)}{\int_{-\infty}^{\infty} L(x_1....x_N, \lambda)\, d\lambda} \tag{116}$$

Instead of confidence limits, probability limits can be given (fig. 22). The likelihood function is not always normalizable. Then a possible way out is a transformation to another parameter with the desired property.

Most physicists violently protest against the Bayesian scheme, even though in practice they use the likelihood recipe to compute confidence limits (descent by 1/2, 2, 4.5.... from the

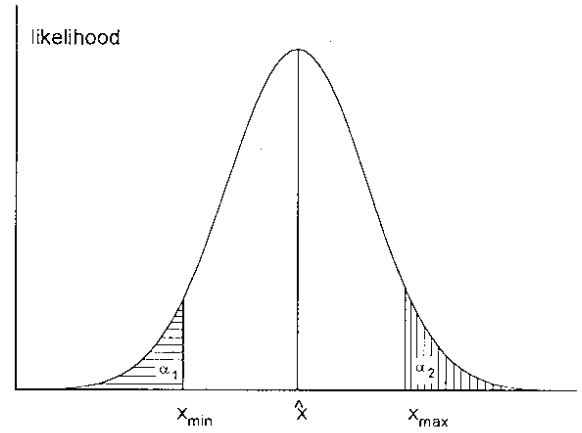

Figure 22: Definition of likelihood limits.

maximum of ln $L$to deduce the 1, 2, 3, .. standard deviation limits) thus ignoring the dogmas of classical statistics.

There are (at least) two cases where the classical and the Bayesian limits agree: When $\hat{x}$ follows a Gaussian of fixed width or when an upper limit is computed for Poisson distributed measurements.

6.3 Complications

6.3.1 Unphysical continuous parameters

In practice it may happen that a measurement or the naively computed confidence limits lye in unphysical regions.

Let us assume that for the measurement of the mass of a particle a value m_e is found with a Gaussian resolution which is independent of its value. This means, that the measurement could also lead to unphysical negative masses or/and that one or both confidence bounds are negative.

Since the confidence limits are unsatisfactory in this case some recipes have been invented to convert the measurement into reasonable limits.

For the special case where the measured or estimated quantity $\hat{x}_0$ follows a Gaussian with fixed variance σ and unknown mean x_{true} the Bayesian approach is the most popular one. (It gives the same results as the classical treatment in cases where no unallowed regions are involved). The normalized likelihood of x_{true}, which is a Gaussian of the same width σ and mean $\hat{x}_0$. The likelihood is normalized in the physical region.

Some papers [32, 33] try to sell this procedure as a "classical" one. Ref. [33] states that the limits obtained are conservative in the sense that they underestimate the confidence. The arguments given, (including my own [32]) are not very convincing.

The general case is more complicated. We will consider three alternative "solutions":

- Through a transformation of variable, $\hat{x}$ is converted into a parameter $\hat{y}$ with Gaussian distribution. Then the standard recipe is applied to find limits for y_{true} which are then transformed back to limits for x_{true}.

- A probability distribution $g(x_{true}; \hat{x})$ ($\hat{x}$ is a parameter) is defined through

$$g(x_{true}; \hat{x}) = \frac{d}{d x_{true}} P(\hat{x}_0 < x_{true}) \quad (117)$$

$$g(x_{true}; \hat{x}) = 0 \quad \text{for } x_{true} \text{ unphysical} \quad (118)$$

  and renormalized to the physical region. Then confidence limits can be computed from the normalized density g_N

$$\int_{-\infty}^{x_{min}} g_N(x_{true}) d\, x_{true} = \alpha_1 \quad (119)$$

$$\int_{x_{max}}^{\infty} g_N(x_{true}) d\, x_{true} = \alpha_2 \quad (120)$$

  Here the lower limit is of little interest. In Figure 23 we give an example. Lines of constant $P(\hat{x}_0 < x_{true})$ are shown. For the observation $\hat{x}$ the 80 % confidence limits are normally given by the values 0.1 and 0.9 at the abscissa. Since the x values below the 0.5 mark are unphysical, the probabilities have to be multiplied by a factor 2. The less stringent limits given in brackets are obtained.

  This method gives the classical results, when there is no forbidden x region.

- The Bayesian method is used, i.e. the likelihood is normalized to the physical region, as shown in Figure 24. The likelihood estimator should be chosen such that it is sufficient and unbiased.

The third suggestion is by far the simplest. Ojectivists probably will prefer the second solution.

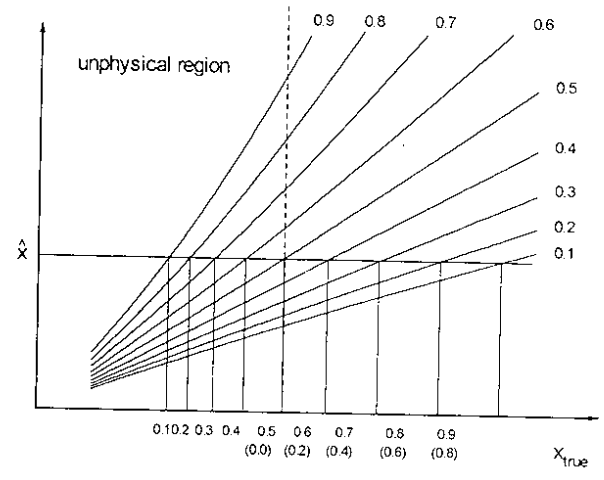

Figure 23: Confidence bounds for unphysical estimates (see text).

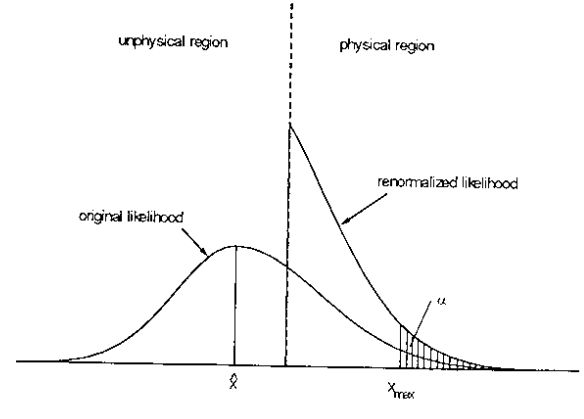

Figure 24: Likelihood limits after renormalization to the physical region

6.3.2 Poisson upper limits in experiments with background

If the outcome of an experiment searching for rare reactions is n events, a $1 - \alpha$ upper confidence limit is found from the relation

$$\sum_{i=0}^{n} P(i; \lambda_{max}) = \alpha \tag{121}$$

with $P(i; \lambda)$ the Poisson distribution for expectation λ. The same limit λ_{max} is obtained in the Bayesian approach:

$$f(\lambda) = \frac{P(n; \lambda)}{\int_0^\infty P(n; \lambda) d\lambda} \tag{122}$$

$$1 - \alpha = \int_0^{\lambda_{max}} f(\lambda) d\lambda \tag{123}$$

When the experimental data are contaminated with background with expectation b an upper limit can be obtained in the Bayesian philosophy [31] and in the classical approach [32]. Again both results agree.

$$\sum_{i=0}^{n} P(i; \lambda_{max} + b) / \sum_{i=0}^{n} P(i; b) = \alpha \tag{124}$$

The result is correct also when the number of observed events is much smaller than the expected background. A treatment is also possible when the background estimate has an uncertainty [32].

6.3.3 Confidence limits for a sample of measurements

Let us consider a series of measurements of the position x_{true} of the center of a slit of known width d (Fig. 25). Each single measurement will give a result $\hat{x}_i$. The observations will follow a uniform distribution of width d and center x_{true}. How can we compute a confidence limit?

There is no obvious solution in the classical picture. One possibility is to combine the individual measurements to a mean value and to compute the confidence limits using the mean. The result of this procedure, however, is not very satisfactory.

Obviously only the extreme measurements $\hat{x}_{min}$ and $\hat{x}_{max}$ carry information and the 100 % confidence limits are clearly $x_{min} \geq \hat{x}_{max} - d/2$ and $x_{max} \leq \hat{x}_{min} + d/2$

The Bayesian method provides these limits (see Fig. 25) and gives also reasonable confidence limits for arbitrary probabilities. A more important and instructive, but also more complex example is the measurement of very short lifetimes through the impact parameter method. The observations may be negative due to the finite resolution and the distribution may be far from exponential. Again it seems very difficult to us to conceive a way to compute sensible classical lower and upper confidence limits. In contrast, the Bayesian method, using the

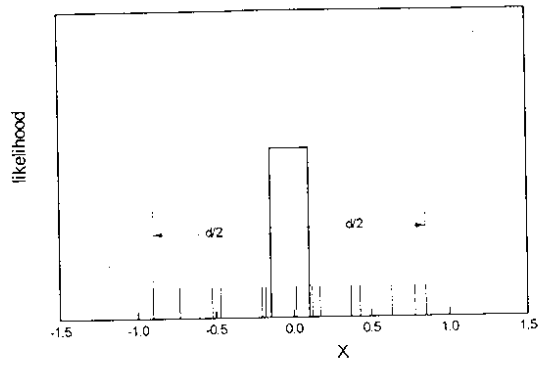

Figure 25: Likelihood limits for the mean of a uniform distribution of width d for a sample of measurements.

normalized likelihood function for the decay constant $\lambda = 1/\tau$ is very attractive and gives reasonable results. However, when we apply the same procedure to the mean life τ we get a surprise: The likelihood function is not normalizable, for large values of τ it behaves like $L(\tau) \propto 1/\tau$.

What should we conclude from this result? Well, Bayesian probabilities depend on the prior probability of the selected parameter. The prior is assumed flat in the normalized likelihood method. Clearly it makes a big difference whether we assume τ or λ to be uniformly distributed. The former assumption is ridiculous: Before the measurement is done nobody would believe that lifetimes between zero and 10^{-10} seconds are by a factor 10^{10} less probable than lifetimes between 1 and 2 seconds! Thus limits for the decay constant λ should be given.Even though the choice of the variable (or the prior) is somewhat arbitrary the result is sensible and well documented.

6.4 Monte Carlo correction

Let us now assume that we compute the mass of a particle from a set of events, where we measure energy and momentum with a certain resolution and a mass dependent acceptance. We would simulate the experiment and fit the mass as described in Section 4 and associate to it the one standard deviation error interval deduced from a decrease of the χ^2 by one or of $\ln L$ by one half.

Such an error interval does not obey the definition of a confidence interval.

A correct way to proceed would be the following: The masses observed in the experiment are averaged to an observed mean value $\overline{m}'$. The experiment is then simulated many times with a certain true value m_{MC}, leading to a distribution of observed mean values $\overline{m}'_{MC}$. To find the lower interval limit, we would modify m_{MC} until the fraction of Monte Carlo experiments with $\overline{m}'_{MC} > \overline{m}'$ is equal to α_1. Correspondingly we could obtain the upper interval limit.

Even with sophisticated weighting techniques this procedure will in most cases require excessive computer time: To get the distribution of mean values $\overline{m}'$ to a precision allowing to determine say a 95 % limit one has to simulate the experiment a few hundred times. If in addition the shape of this distribution varies with x_{true} one has to repeat the whole procedure several times.

(The computation of upper limits for Poisson distributed numbers only requires an acceptance correction, with can easily computed with a standard Monte Carlo simulation.)

6.5 A plea for the use of likelihood limits

The concept of confidence limits has been invented to quantify the precision of a measurement. Any subjective assumptions made by the research person are carefully avoided.

The concept works well for Poisson distributed measurements. There confidence limits are equivalent to likelihood limits.

There are several unsolved problems with confidence limits for continuous measurements or estimates, due to:

- external bounds to the variables from laws of physics,
- the necessity to handle samples of measurements,
- the need to apply extensive Monte Carlo corrections,
- the need to combine results from different experiments,
- the need to compute limits for one parameter from fits to several parameters.

We would like to advocate for replacing *confidence limits* by *likelihood limits* which avoid all of those problems.

The main objection to this Bayesian approach is to the introduction of subjective elements into physics, which is considered as an "act of desperation" [6]. Also the association of a probability density to a fixed quantity like a particle mass or lifetime is objected.

In deed the original justification of likelihood limit method is based on Bayes theorem with the assumption of an uniform prior of the likelihood parameter. A different choice of the

prior distribution is equivalent to a different choice of the parameter. Thus choosing an uniform prior still leaves full freedom to us. The only subjective element is then the choice of the likelihood variable. We could in addition fix the likelihood variable by the requirement that the likelihood function be Gaussian. However this is not really necessary (but likelihood functions with long tails should be avoided), if the variable used is documented.

The problems cited above are all absent with likelihood limits. Unphysical regions are taken into account by a correct normalization of the likelihood, no special treatment is necessary for samples of measurements, and Monte Carlo corrections are done as explained in section 4.

Results of different experiments can be combined by adding the log-likelihoods. These, or enough information to reconstruct them should be given in the publications.

It will not be easy to convince physicists, that Bayesian methods provide useful concepts and are objective in the sense that the procedures used can be well defined. An interesting discussion of classical versus Bayesian confidence limits can be found in a recent publication [34]. The conclusions are similar to ours but stick to classical results for certain cases.

A Concept of 'equivalent number of events'

Example: In a cross section measurement half (10 events) of the total event sample (20 events) are observed with acceptance 1 , the other half with acceptance 0.5. We combine 10 events with weight $w_1 = 1$ and 10 events with weight $w_2 = 2$ to the weighted sum k of 30. The estimated absolute error is $\delta k = \sqrt{10w_1^2 + 10w_2^2} = \sqrt{50}$ the relative error $\delta k/k = 1/\sqrt{(18)}$. Thus the 20 weighted events have the same statistical significance as 18 unweighted events. The "number of equivalent events" is 18.

We now generalize and extend this concept.

Be k the weighted sum of N Poisson distributed numbers k_i with weights w_i and mean values *one*.

$$k = \sum_{i=1}^{N} w_i k_i \tag{125}$$

If all weights were equal to *one*, the distribution $W(k)$ of k were a Poisson distribution with mean N. In the general case the mean and variance of the discrete distribution $W(k)$ are

$$E(k) = \sum_{i=1}^{N} w_i \tag{126}$$

$$Var(k) = \sum_{i=1}^{N} w_i^2 \tag{127}$$

For χ^2 estimates it is sufficient to know mean and variance of $W(k)$ but with low event numbers likelihood analysis are necessary and then $W(k)$ itself has to be known to reasonable precision. It is a complicated distribution, however it can be approximated by a Poisson distribution using the concept of *equivalent number of events*. To this end we compute a number $\tilde{k}$ with variance $\tilde{k}$ which has the same relative error as k. Thus $\tilde{k}$ unweighted events have the same statistical significance as the N weighted events.

We set

$$\tilde{k} = \frac{(\sum_{i=1}^{N} w_i k_i)^2}{\sum_{i=1}^{N} w_i^2 k_i} \tag{128}$$

and obtain

$$E(\tilde{k}) = \frac{(\sum_{i=1}^{N} w_i)^2}{\sum_{i=1}^{N} w_i^2} = Var(\tilde{k}) \tag{129}$$

The relative scale s between $E(k)$ and $E(\tilde{k})$ is

$$s = \frac{E(k)}{E(\tilde{k})} = \frac{\sum_{i=1}^{N} w_i^2}{\sum_{i=1}^{N} w_i} \tag{130}$$

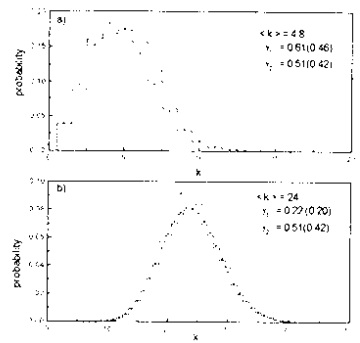

Figure 26: Comparison between Poisson distribution and distribution of equivalent number of events, consisting of a mixture of events with weight one and ten. The values of γ_1 and γ_2 for the Poisson distribution are given in parentesis.

By construction we now have $E(k/s) = E(\tilde{k})$ and $Var(k/s) = Var(\tilde{k})$. An interesting question is, whether it is possible to compute also the higher moments of the distribution of k/s from the Poisson distribution. Since it is hard to solve this problem analytically, we try to obtain a qualitative result through a Monte Carlo simulation.

In Figure 26 we present two examples which illustrate the similarity between the distribution of k/s and the Poisson distribution of $\tilde{k}$. In both examples we chose an equal number of events with weight 1 and 10, respectively, which corresponds to a scale factor of about 9.18. One weighted event has a statistical significance of about 0.60 Poisson distributed events.

To compare the two discrete distributions we integrated them over one unit (which leaves the Poisson distribution numerically unchanged). The resulting histograms are very similar in both cases, the agreement being better in the case with the higher number of equivalent events. The Poisson distribution describes the weighted sum the better, the higher is the equivalent number of events and the smaller is the difference between the weights. The coefficients γ_1 of the skewness and of the kurtosis γ_2 computed from the Poisson distributions in our examples are in reasonable agreement with those of the weighted distributions, which are about 20 % higher.

In Figure 27 we simulate background subtraction. An average of 20 events with weight 1 and of 4 events with weight -1 are generated. The statistical significance corresponds to 10.7 unweighted events without background. The fluctuations in the histogram of Figure 27 is an artifact of the binning.

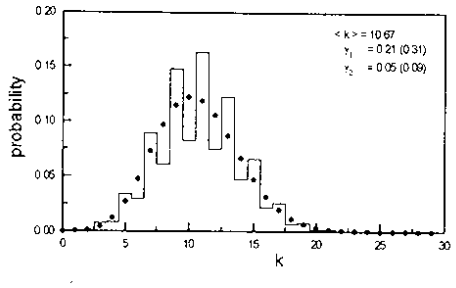

Figure 27: Comparison between Poisson distribution and distribution of equivalent number of events, consisting of a mixture of events with weight one and 4 background events with weight -1.

The concept of equivalent Poisson distributed events is readily extended to a continuous distribution of weights, replacing the sums in the above relations by integrals. For a sample of Poisson distributed events with mean N and weights distributed according to the probability density $g(w)$ of w we obtain:

$$E(k) = N \int_{-\infty}^{\infty} g(w)dw \tag{131}$$

$$Var(k) = N \int_{-\infty}^{\infty} g(w)w^2 dw \tag{132}$$

$$\tilde{k} = N \frac{\left(\int_{-\infty}^{\infty} g(w)dw\right)^2}{\int_{-\infty}^{\infty} g(w)w^2 dw} \tag{133}$$

As an example we have generated event according to a Poisson distribution with mean 20 and selected weights following a uniform distribution between 0 and 1. The number of equivalent events is 15. Figure 28 illustrates the fairly good agreement of the two distributions.

In practice the a priori distribution of the weights is often not known. One is faced with a sample of weighted events $\{w_1, w_2, ...w_N\}$. The the equivalent number of events is estimated using relation 129.

$$\tilde{k} \approx \frac{(\sum_{i=1}^{N} w_i)^2}{\sum_{i=1}^{N} w_i^2} \tag{134}$$

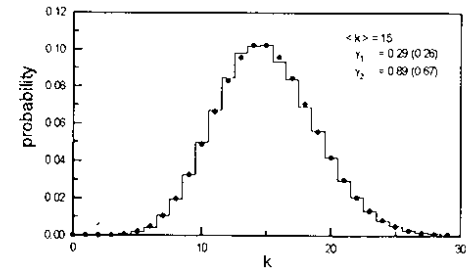

Figure 28: Comparison between Poisson distribution and distribution of equivalent number of events, consisting of a mixture of events with continuous weights between zero and one.

B Minimum detectable systematic error in a χ^2 test

The statistical variance of the χ^2-distribution is, for not too small a number B of bins equal to $2B$. Thus we roughly measure the χ^2-value to a precision of $\sqrt{2B}$ and we may argue that we can detect a systematic error, contributing $\sqrt{2B}$ to χ^2. For instance for systematic deviations α_μ we would expect

$$\begin{aligned} < \chi^2 > &= \sum_\mu \frac{< (d'_\mu + \alpha_\mu - m'_\mu)^2 >}{m'_\mu} \\ &= B + \sum_\mu \frac{< \alpha_\mu^2 >}{m'_\mu} \\ &= B(1 + < \frac{\alpha_\mu^2}{m'_\mu} >) \end{aligned}$$

Now we assume that the systematic deviation is proportional to the number of events in a bin $\alpha_\mu \approx \alpha_0 m'_\mu$ (think for instance of a constant fraction of background). We set $B < m'_\mu = N$ where N is the total number of events and get

$$< \chi^2 >= B + \alpha_0^2 N$$

Following our argument we can detect the additional contribution if it exceeds $\sqrt{2B}$ and find for the minimum detectable deviation

$$\alpha_0 \approx \frac{(2B)^{1/4}}{N^{1/2}}$$

The ratio of the detectable systematic error to the statistical error is inversely proportional to the fourth root of the number of bins.

C Computing EDF test probabilities

The summary given here follows an article by M. A. Stephens [10]. The results given there are extended to account for the prediction to be given through a Monte Carlo simulation.

Computation of test statistics

The computation of supremum statistic D is straight forward.

The quadratic statistics W^2 and A^2 are computed after a Probability Integral Transformation (PIT). The PIT transforms the expected theoretical distribution of x_{MC} into a uniform distribution of z_{MC}. The variable z is obtained from the relation $z = F(x)$ where F is the cumulative distribution of x.

We know the distribution $F(x)$ only indirectly through the Monte Carlo simulation. The value Z for an experimental event located at X is roughly $Z \approx$ (# of MC events with X_{MC} ¡ X) /(total # of MC events). A slightly more precise estimate is obtained by interpolation.

We obtain W^2 and A^2 from

$$W^2 = \frac{1}{12N} + \sum_i (Z_i - \frac{2i-1}{2N})^2 \tag{135}$$

$$A^2 = -N + \sum_i (Z_i - 1)\,(\ln Z_i + \ln(1 - Z_{N+1\ -i})) \tag{136}$$

Computation of probability

For a large 'number of events' N, there exist modified test statistics, D^*, W^{2*}, A^{2*} which are, for large N, independent of N. They are defined by the following empirical relations, which are good approximations for $N > 20$.

$$D* = D_{max}(\sqrt{N} + 0.12 + \frac{0.11}{\sqrt{N}}) \tag{137}$$

$$W^2* = (W^2 - \frac{0.1}{N} + \frac{0.6}{N^2})(1.0 + \frac{1.0}{N}) \tag{138}$$

$$A^2* = A^2 \tag{139}$$

The relation between the modified test statistics and the test probability are given in Figure 29.

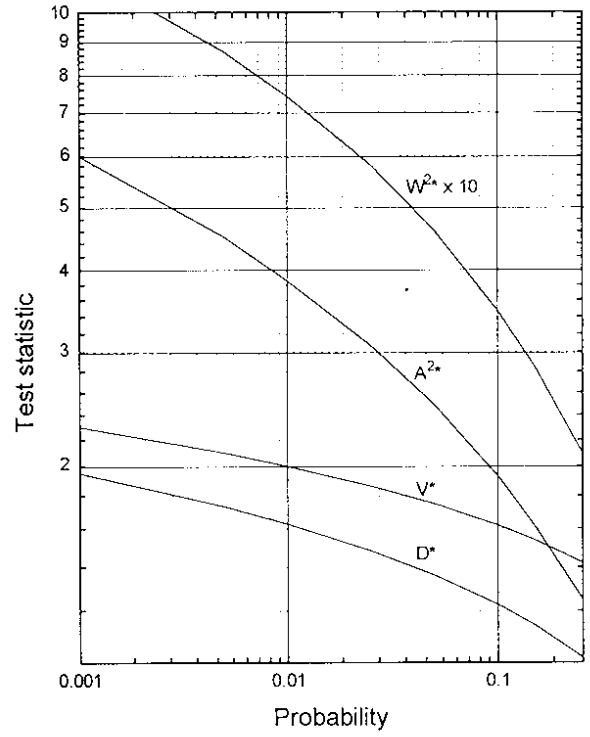

Figure 29: Probabilities for modified EDF statistics

D Likelihood comparison of experimental data with simulation

Likelihood for a Poisson distributed number from a sample of two elements

If we have a sample {n,m} of two elements from a Poisson distribution the logarithmic likelihood for its mean θ is

$$\ln L = n \ln \theta - \theta + m \ln \theta - \theta - \ln n! - \ln m! + const. \tag{140}$$

In a maximum likelihood analysis we set the derivative of $\ln L$ with respect to θ equal to zero. The constant term containing the factorials is without interest and can be omitted.

$$\frac{\partial \ln L}{\partial \theta} = 0 \quad \longrightarrow \quad \hat{\theta} = \frac{n+m}{2} \tag{141}$$

Now we assume that the n and m correspond to two measurements, made under different conditions (for example different acceptances) such that the expectation for n is $c_n\theta$ and for m is $c_m\theta$. The likelihood becomes

$$\ln L = n \ln c_n\theta - c_n\theta + m \ln c_m\theta - c_m\theta + const. \tag{142}$$

and the value of θ which maximizes L is

$$\hat{\theta} = \frac{n+m}{c_n + c_m} \tag{143}$$

Comparison of experimental data with parameter dependent simulation

Now we compare a bin content of an experimental histogram with a Monte Carlo prediction. Both the data and the prediction be Poisson distributed up to a scaling factor. To be precise, we have $\tilde{d}$ and $\tilde{m}$ equivalent events and d and m weighted or scaled events. Then the log-likelihood for the true expectation θ to be compatible with both the data and the simulation is according to (142):

$$\ln L = \tilde{d} \ln \frac{\tilde{d}}{d}\theta - \frac{\tilde{d}}{d}\theta + \tilde{m} \ln \frac{\tilde{m}}{m}\theta - \frac{\tilde{m}}{m}\theta + const. \tag{144}$$

Here we neglect the fluctuations in the scaling factors. The best estimate for θ is then

$$\hat{\theta} = \frac{\tilde{d} + \tilde{m}}{\frac{\tilde{d}}{d} + \frac{\tilde{m}}{m}} \tag{145}$$

which, inserted into the likelihood function gives

$$\ln L = -(\tilde{d} + \tilde{m}) \ln(\frac{\tilde{d}}{d} + \frac{\tilde{m}}{m}) - \tilde{d} \ln d - \tilde{m} \ln m + const. \tag{146}$$

Terms containing only $\tilde{d}$ and $\tilde{m}$ have been absorbed in the constant.

Let us assume the Monte Carlo prediction m depends on a parameter λ which we are interested in. We than have to minimize $\ln L$ with respect to λ. Now also terms depending only on d can be ignored.

$$\ln L = -(\tilde{d} + \tilde{m}) \ln(\frac{\tilde{d}}{d} + \frac{\tilde{m}}{m}) \quad \tilde{m} \ln m + const. \tag{147}$$

To get the maximum we do not derive (147) with respect to λ, we would get a trivial result. In fact we first have to sum all histogram bins and then search for the maximum of the sum (see Section 4.2)

Acknowledgement

I would like to thank Prof. V. Blobel for interesting discussions, detailed explanations on his unfolding method and very valuable comments to this report.

References

[1] W. T. Eadie et al., Statistical Methods in Experimental Physics, Noth-Holland, Amsterdam (1971)

[2] S. Brandt, Statistical and Computational Methods in Data Analysis, North-Holland, Amsterdam, 1976

[3] A. G. Frodesen, O. Skjeggestad, M. Tofte, Probability and Statistics in Particle Physics, Universitetsforlaget, Bergen, 1979

[4] T. A. Bancroft and C.-P. Han, Statistical Theory and Inference in Research, M. Dekker, Inc. New York (1981)

[5] L. Lyons, Statistics for nuclear and particle physicists, Cambridge University Press, 1992

[6] B. P. Roe, Probability and Statistics in Experimental Physics, Springer-Verlag, New York, 1992

[7] R. B. d'Agostino and M. A. Stephens (editors), Goodness of Fit Techniques, M. Dekker, New York (1986)

[8] D. S. Moore, Tests of Chi-Squared Type, in Goodness of Fit Techniques, ed. R. B. d'Agostino and M. A. Stephens, M. Dekker, New York (1986)

[9] M. A. Stephens, Tests based on EDF Statistics in Goodness of Fit Techniques, ed. R. B. d'Agostino and M. A. Stephens, M. Dekker, New York (1986)

[10] M. A. Stephens, Tests for the Uniform Distribution in Goodness of Fit Techniques, ed. R. B. d'Agostino and M. A. Stephens, M. Dekker, New York (1986)

[11] H. B. Mann and A. Wald, Ann. Math. Statist. 13 (1942) 306

[12] MINUIT Function minimization, F. James and M. Roos, CERN note D506 Nucl. Instr. and Meth. A340 (1994) 396

[13] P. Eberhard, G. Lynch and D. Lambert, Fits of Monte Carlo distributions to data, Nucl. Instr. and Meth. A326 (1993) 573

[14] G. Zech et al., A Measurement of the Lifetimes of Ξ^0 and Λ Hyperons, Nucl. Phys. B124 (1977) 413

[15] G. Zech, A Monte Carlo Method for the Analysis of Low Statistic Experiments, Nucl. Instr. and Meth. 137 (1978) 551.

[16] F. James, Monte Carlo theory and practice, in: Experimental techniques in high energy physics, ed. T.Ferbel, Addison-Wesley (1987), 627

[17] D. M. Schmidt, R. J. Morrison and M. S. Witherell, A general method of estimating physical parameters with acceptance and smearing effects, Nucl. Instr. and Meth. A328 (1993) 547

[18] V. Blobel, Unfolding methods in high-energy physics experiments, Proc. 1984 CERN School of Computing, Aiguablava, Spain, CERN 85-09 (1984) 88.

[19] V. B. Anykeyev, A. A. Spiridonov and V. P. Zhigunov, Comparative investigation of unfolding methods, Nucl. Instr. and Meth. A303 (1991) 350.

[20] L. A. Shepp and Y. Vardi, IEEE trans. Med. Imaging MI-1 (1982) 113.

[21] A. Kondor, Method of converging weights - an iterative procedure for solving Fredholm's integral equations of the first kind. Nucl. Instr. and Meth. 216 (1983) 177

[22] H. N. Müllthei and B. Schorr, On an iterative method for the unfolding of spectra, Nucl. Instr. and Meth. A257 (1986) 371

[23] G. I. Marchuk, Methods of Numerical Mathematics, Springer, Berlin (1975).

[24] J. D'Agostini, A Multidimensional unfolding method using Bayes theorem, Nucl. Instr. and Meth.(1994).

[25] M. Schmelling, The method of reduced cross-entropy - a general approach to unfold probability distributions, Nucl. Instr. and Meth. A340 (1994) 400

[26] L. Lindemann and G. Zech, Unfolding by weighting Monte Carlo events, Nucl. Instr. and Meth. A354 (1994) 516

[27] D. L. Lindley, The 1988 Wald Memorial Lectures: The present Position in Bayesian Statistics, Statistical Sci. 5,1 (1990) 44

[28] F. James and M. Roos, Statistical notes on the problem of experimental observations near an unphysical region, Phys. Rev. D44 (1991) 299

[29] A. A. Marchetti, Deconvolution of mass spectra, Nucl. Instr. and Meth. A324 (1993) 281

[30] 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. Instr. and Meth. A340 (1994) 396

[31] O. Helene, Upper limit of peak area, Nucl. Instr. and Meth. 212 (1983) 319

[32] G. Zech, Upper limits in experiments with background or measurement errors, Nucl. Instr. and Meth. A340 (1994) 396

[33] K. Hikasa et al., Review of particle properties, Phys. Rev. D45 (1992)

[34] D. Cousins, Why Isn't every physicist a Bayesian?, UCLA-HEP-94-5 (1994).