DEUTSCHES ELEKTRONEN-SYNCHROTRON

DESY

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.