ELSEVIER

Nuclear Instruments and Methods in Physics Research A 425 (1999) 357–360

NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A

Linear interpolation of histograms

A.L. Read[1]

University of Oslo, Department of Physics, P.O. Box 1048, Blindern, 0316 Oslo, Norway

Received 19 October 1998

Abstract

A prescription is defined for the interpolation of probability distributions that are assumed to have a linear dependence on a parameter of the distributions. The distributions may be in the form of continuous functions or histograms. The prescription is based on the weighted mean of the inverses of the cumulative distributions between which the interpolation is made. The result is particularly elegant for a certain class of distributions, including the normal and exponential distributions, and is useful for the interpolation of Monte Carlo simulation results which are time-consuming to obtain.

PACS: 07.05.K; 07.05.T

Keywords: Analysis; Distribution; Histogram; Interpolation; Simulation

1. Introduction

In a typical new particle search performed with a high-energy physics experiment, Monte Carlo simulations of the detector response to the hypothetical particle may be performed on a mass (or other model parameter) grid, e.g. the reconstructed invariant mass (or other event variable) distribution of the hypothetical candidates is produced at certain masses, while the object of the search is to either discover or exclude the signal at an arbitrary mass. In order to compute signal and background confidence levels at arbitrary masses it is necessary to do some sort of interpolation between the mass points where Monte Carlo simulation results are available. If the mass resolution of the detector is finer than the step-size of the grid of simulation results, a simple interpolation of the measured confidence levels may be a poor approximation. The question is how to take into account kinematic suppressions, experimental cutoffs, evolutions of resolution, i.e., all the features of the distribution which evolve with the model parameter in a complicated analysis.

2. Distribution interpolation defined

The method proposed here is based on the linear interpolation of the inverses of the cumulative distribution functions (c.d.f.s). Two probability

[1] Tel.: +47 22 855062; fax: +47 22 856422 ; e-mail: alex.read@fys.uio.no.

0168-9002/99/$ – see front matter
PII: S0168-9002(98)01347-3

distribution functions (p.d.f.s) for the observable x, $f_1(x)$ and $f_2(x)$ (defined in the physical region $x \geq -\infty$), have corresponding c.d.f.s

$$F_1(x) = \int_{-\infty}^{x} f_1(x')\,\mathrm{d}x', \tag{1}$$

$$F_2(x) = \int_{-\infty}^{x} f_2(x')\,\mathrm{d}x'. \tag{2}$$

The goal is to obtain a new p.d.f. $\bar{f}(x)$ with its corresponding c.d.f.

$$\bar{F}(x) = \int_{-\infty}^{x} \bar{f}(x')\,\mathrm{d}x' \tag{3}$$

that describe the new distribution. The first step of the interpolation procedure is to find x_1 and x_2 where the cumulative distributions F_1 and F_2 are equal for a given cumulative probability y,

$$F_1(x_1) = F_2(x_2) = y. \tag{4}$$

The cumulative probability for the new distribution $\bar{F}$ is set to the same value y at a linearly interpolated position x,

$$\bar{F}(x) = y, \tag{5}$$

$$x = ax_1 + bx_2. \tag{6}$$

The constants a and b express the interpolation distance between the extreme values of the relevant parameter for the two existing distributions (and by construction satisfy $a + b = 1$). For a new particle search, for instance, this could be the relative position of the mass hypothesis between two masses for which the observable invariant mass distributions have been computed in a Monte Carlo simulation.

The p.d.f. $\bar{f}(x)$ is obtained by inverting the cumulative distributions in Eqs. (4) and (5), substituting these results in Eq. (6),

$$\bar{F}^{-1}(y) = aF_1^{-1}(y) + bF_2^{-1}(y), \tag{7}$$

deriving this with respect to y and solving for the interpolated p.d.f. $\bar{f}(x)$,

$$\bar{f}(x) = \frac{f_1(x_1)f_2(x_2)}{af_2(x_2) + bf_1(x_1)}. \tag{8}$$

3. Interpolating distributions of the form $g((x-x_o)/k)$

The prescription for linear interpolation of p.d.f.s defined in Eqs. (4)–(6) takes a particularly elegant form when the initial p.d.f.s can both be described by the same function with different arguments of a particular form. Suppose that

$$f_1(x) = g\left(\frac{x - x_{1o}}{k_1}\right), \tag{9}$$

$$f_2(x) = g\left(\frac{x - x_{2o}}{k_2}\right), \tag{10}$$

$$G\left(\frac{x - x_{io}}{k_i}\right) = \int_{-\infty}^{x} g\left(\frac{x' - x_{io}}{k_i}\right)\mathrm{d}x', \tag{11}$$

where x_{io} and k_i regulate the offset and scale features of p.d.f. i. Eq. (4) takes the form

$$G\left(\frac{x_1 - x_{1o}}{k_1}\right) = G\left(\frac{x_2 - x_{2o}}{k_2}\right) = y. \tag{12}$$

This implies that the arguments in Eq. (12) have the same value λ,

$$\lambda = \frac{x_1 - x_{1o}}{k_1} = \frac{x_2 - x_{2o}}{k_2}. \tag{13}$$

Substitution of these results in Eq. (5) yields

$$\bar{F}(x) = G(\lambda), \tag{14}$$

i.e. the same function describes both the initial and the interpolated cumulative distributions. Substitution in Eq. (6) yields

$$\lambda = \frac{x - \bar{x}_0}{\bar{k}}, \tag{15}$$

where

$$\bar{x}_0 = ax_{1o} + bx_{2o}, \tag{16}$$

$$\bar{k} = ak_1 + bk_2. \tag{17}$$

The interpolated p.d.f. $\bar{f}(x)$ is then obtained by derivation of $G(\lambda)$,

$$\bar{f}(x) = g\left(\frac{x - \bar{x}_0}{\bar{k}}\right). \tag{18}$$

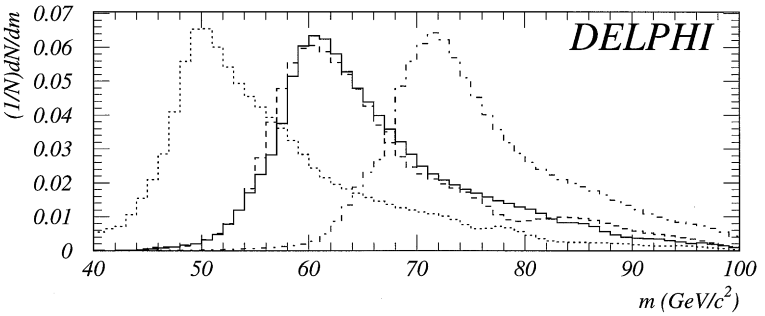

Fig. 1. Example of histogram interpolation. The reconstructed invariant mass distribution for a 60 GeV/c^2 Higgs (solid line) has been obtained by interpolation between the simulation results for 50 (dotted line) and 70 (dash-dotted line) GeV/c^2 Higgs masses and is overlaid with the simulation results of a 60 GeV/c^2 Higgs (dashed line).

This means, for instance, that if both initial p.d.f.s are normal distributions, the interpolated distribution is also a normal distribution with both the variance and mean linearly interpolated between those of the initial distributions. For exponential decay distributions ($x_0 = 0, k < 0$) the decay constant of the interpolated distribution is linearly interpolated between the decay constants of the initial distributions. The same elegant result is also obtained whenever both initial distributions can be described by the same power series in $(x - x_{io})/k_i$.

4. Interpolation of histograms

For a pair of initial distributions obtained by histogramming the output of a complex physics and detector simulation, it may not be practical to identify a particular function which can describe both initial distributions, but it is always possible to carry out the computation in Eqs. (4)–(8) on the histograms directly in order to obtain an interpolated histogram. Fig. 1 shows how well the interpolation procedure manages to describe the reconstructed invariant mass distribution of a hypothetical Standard Model Higgs boson with mass $m_H = 60\,\mathrm{GeV}/c^2$ decaying in the $ZH^0 \to e^+e^-H^0$ channel at center-of-mass energy $\sqrt{s} = 161\,\mathrm{GeV}$ based on simulations at $m_H = 50$ and $70\,\mathrm{GeV}/c^2$ [1], despite the large differences between the 3 mass distributions (these distributions were chosen to illustrate the interpolation precisely because they have several, rapidly evolving features which could be difficult to interpolate). Since the interpolation is able to predict so well the main features of the $60\,\mathrm{GeV}/c^2$ simulation results, it is plausible that the smaller interpolation distances ($\leq 2.5\,\mathrm{GeV}/c^2$) used in the DELPHI Higgs search will lead to results superior to those obtained by more simple means (for example simply using the nearest available distribution or shifting it by the difference between the mass hypothesis and the simulated mass).

5. Conclusion

A method for linearly interpolating probability distribution functions or histograms has been presented here. The method gives particularly intuitive results for certain well-known probability distribution functions, such as the normal distribution, where the mean and variance parameters are simply linearly interpolated between those of the initial pair of distributions. The method can also be applied to histogrammed simulation data where the distribution functions are not known and are very different from each other, and surprisingly reasonable results may be obtained.

Acknowledgements

This work was supported by a grant from the Norwegian Research Council. I would like to thank my DELPHI colleagues for permission to use their published Higgs search simulation results to illustrate the histogram interpolation.

References

[1] DELPHI Collaboration, Search for neutral and charged Higgs bosons in e^+e^- collisions at $\sqrt{s} = 161\,\text{GeV}$ and 172 GeV, Eur. Phys. J. C 2 (1998) 1.