Elements of statistical methods in High Energy Physics analyses

Bertrand Laforge, Laurent Schoeffel

Service de Physique des Particules
DAPNIA, CEA Saclay

Abstract

The present document intends to give examples of statistical treatement in physics analyses, we recall definition of very useful quantities such as acceptance, purity and efficiencies then we derive error calculations for these quantities for weighted or unweighted Monte Carlo as well as for data. We give also in this discussion a few basic ideas and theorems needed in the statistical treatement of random variables and we examplify the use of these theorems in the error computations for the different physical quantities we are studying here. The set of formulae we derive can find a direct application in the analyses of the proton structure functions and related subjects in the H1 experiment, these results are also well adapted to determine the statistical limits of a given analysis.

1 Introduction

In the following we recall some basic definitions for useful quantities such as acceptance, purity and efficencies and we present a rigorous statistical treatment for them. We insist on the care that should be taken in defining independant variables and then we show how to compute error calculations for physical quantities as functions of these variables. This is in practice important when we want to identify some statistical limits for a given analysis [1, 2]. We end our discussion with a few comments on χ^2 calculations for a set of random variables which can give a better understanding of probability distributions. These calculations can find a direct application in the analyses of the proton structure functions and related subjects developped in the H1 experiment.

2 Principles of error calculation

2.1 Formula for error propagation

Let f be a function of the n variables $\{x_i\}$. The variables $\{x_i\}$ can take any number of different values as a result of a measurement, we call them random variables. The probability associated with measuring each of the possible values of the $\{x_i\}$ variables form a probability distribution.

$$f = f(x_1, \ldots, x_n)$$

If the n variables are uncorrelated then the error on the value taken by f can be expressed as a function of the errors on the variables $\{x_i\}$ through the following formula [1]

$$\sigma_f^2 = \sum_i \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_{x_i}^2 \tag{1}$$

It is very important to remind that formula is no longer valid if the n variables $\{x_i\}$ are correlated. Thus error calculations using formula (1) impose to identify uncorrelated (independant) variables. The following computations examplify this point.

[1] When the $\{x_i\}$ are independant variables then the probability distribution f can be written as a product of the probability distributions of all variables

$$f = f(x_1, \ldots, x_n) = f_1(x_1) \ldots f_n(x_n)$$

2.2 Weighted Monte Carlo

For unweighted Monte Carlo, if the number of events in a bin is N then the statistical error on this number is $\sigma_N = \sqrt{N}$. We can give a short proof of this property.

> We note p the probability for an event to be in a bin $\mathcal{B}$. For a total number of n events we have approximately $pn = N$ events in the bin $\mathcal{B}$. The probability for having k events in $\mathcal{B}$ is
>
> $$pr(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$
>
> or if we note $\lambda = np$
>
> $$pr(k) = \frac{n^k (\frac{\lambda}{n})^k (1-\frac{\lambda}{n})^n}{k!}$$
>
> for large n ($n \gg k$) it is staightforward to see that
>
> $$pr(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$
>
> For this distribution we can easily get the corresponding moments
>
> $$E(k) = \lambda$$
>
> $E(k)$ stands for the mean value of k.
>
> $$\sigma^2(k) = \lambda$$
>
> As $N = \lambda$ we can find that in the bin $\mathcal{B}$ we have $N \pm \sqrt{N}$ events.

Thus the contribution of each event to the error squared is $N/N = 1$. Now for weighted Monte Carlo if these events are generated with the same weight W_0 then the content of the bin is $W_0 N$ and the error is $\sigma = W_0\sqrt{N}$ i.e. $\sigma^2 = W_0^2 N$. So each event gives a contribution to the squared error which is $(W_0 N)/N = W_0$. If each event is generated with a weight W_i then the content of the bin is $\sum\limits_i W_i$ and the squared error is $\sum\limits_i W_i^2$. In the next sections we use the notation

$$\sigma_{W_i}^2 = W_i^2 \tag{2}$$

3 Acceptance Computation I

3.1 General case

Acceptance computation consists of evaluating the ratio of events reconstructed in a bin over the one generated in it. We note n the events reconstructed in the bin and p the events generated in the bin, we have for the acceptance

$$A = \frac{N_{rec}}{N_{gen}} = \frac{\sum\limits_n W_n}{\sum\limits_p W_p} = \frac{a}{b}$$

But the two terms $a = \sum\limits_n W_n = N_{rec}$ and $b = \sum\limits_p W_p = N_{gen}$ are not independent since the number of events reconstructed in the bin is a fraction of the number of events generated in the bin. So, we have to identify what terms are independent from the others. There are three groups of independant events, namely the events generated and reconstructed in the bin (denoted i), the events generated out of the bin and reconstructed in the bin (denoted j) and the events generated in the bin and reconstructed outside of the bin (denoted k). Then we have

$$N_{rec} = N_{stay} + N_{come} \qquad\qquad N_{gen} = N_{stay} + N_{leave}$$

and the acceptance is

$$A = \frac{N_{stay} + N_{come}}{N_{stay} + N_{leave}} = \frac{\sum\limits_i W_i + \sum\limits_j W_j}{\sum\limits_i W_i + \sum\limits_k W_k} \tag{3}$$

In this equation $\sum\limits_i W_i = N_{stay}$, $\sum\limits_j W_j = N_{come}$, $\sum\limits_k W_k = N_{leave}$ are independant variables and we can apply formula (1) for this variables to evaluate the error on A. Note that $\sum\limits_i W_i = N_{stay}$, $\sum\limits_j W_j = N_{come}$, $\sum\limits_i W_i + \sum\limits_k W_k = N_{gen}$ are not independant, so the formula (1) can not be applied for them. Then we calculate the error on the acceptance

$$\sigma_A^2 = \sum_i (\frac{\partial A}{\partial W_i})^2 \sigma_{W_i}^2 + \sum_j (\frac{\partial A}{\partial W_j})^2 \sigma_{W_j}^2 + \sum_k (\frac{\partial A}{\partial W_k})^2 \sigma_{W_k}^2$$

where $$\frac{\partial A}{\partial W_i} = \frac{b-a}{b^2} \qquad \frac{\partial A}{\partial W_j} = \frac{1}{b} \qquad \frac{\partial A}{\partial W_k} = \frac{-a}{b^2}$$

and, since according to formula (2) $\sigma^2_{W_\xi} = W^2_\xi$ for $\xi \in [i,j,k]$, the error squared on the acceptance can be written

$$\sigma^2_A = (\frac{b-a}{b^2})^2 \sum_i {W_i}^2 + \frac{1}{b^2} \sum_j {W_j}^2 + (\frac{-a}{b^2})^2 \sum_k {W_k}^2 \tag{4}$$

3.2 Unweighted Case

In the unweighted case, we have $W_\xi = 1.$ with $\xi \in [i,j,k]$.It is straightforward to see that

$$\sigma^2_A = \frac{(N_{stay} + N_{come})^2}{{N_{gen}}^3} (\frac{N_{gen} - 2N_{stay}}{N_{stay} + N_{come}} + 1) \tag{5}$$

We can then derive two interesting inequalities :
if $N_{gen} \leq N_{rec}$ we have

$$(\frac{\sigma_A}{A})^2 \geq (\frac{N_{gen} - N_{rec}}{{N_{rec}}^2}) + \frac{2N_{come}}{{N_{rec}}^2}$$

and if $N_{gen} \geq N_{rec}$ we have

$$(\frac{\sigma_A}{A})^2 \geq \frac{2N_{come}}{{N_{gen}}^2}$$

These relations determine the minimal relative error we can expect on the acceptance when we know the rate of migrations, namely $\frac{N_{come}}{N_{rec}}$, this rate depends of course on the bin $\mathcal{B}$ considered.

4 Acceptance Computation II

We now reproduce acceptance calculations in the case where we change the reference system. We analyse the case for which in the new reference system the values for the different weights ($W_\xi = 1.$ with $\xi \in [i,j,k]$) of our previous analysis are multiplied by a flux factor, this is what happens for example in an $electron - proton$ collision when we pass from the $e-p$ center of mass system to the $\gamma^\star - p$ system of reference. We note ϕ this factor, as ϕ is a function of the kinematic variables we have to distinguish ϕ_{rec} and ϕ_{gen} then we can express N_{rec} and N_{gen}

$$N_{rec} = N_{rec,stay} + N_{rec,come}$$

$$N_{gen} = N_{gen,stay} + N_{gen,leave}$$

We keep the notation of section (3), it means that the index (i) stands for the group of events which stay in the bin considered, the index (j) stands for the group of events which come into the bin considered and the index (k) stands for those which leave the bin considered. We have

$$N_{rec,stay} = \sum_i W_i \phi_{rec,i}$$

$$N_{rec,come} = \sum_j W_j \phi_{rec,j}$$

$$N_{gen,stay} = \sum_i W_i \phi_{gen,i}$$

$$N_{gen,leave} = \sum_k W_k \phi_{gen,k}$$

The definition of the acceptance becomes

$$A = \frac{\sum\limits_i W_i \phi_{rec,i} + \sum\limits_j W_j \phi_{rec,j}}{\sum\limits_i W_i \phi_{gen,i} + \sum\limits_k W_k \phi_{gen,k}} = \frac{a}{b} \qquad (6)$$

The method for deriving error calculation for this quantity is similar to the one we explained in setion (3) but now $\phi_{rec,i}$ and $\phi_{rec,j}$ are also independant variables and we have to take this into account. We obtain by applying formula (1)

$$\begin{aligned} \sigma_A^2 &= \sum_i (\frac{\partial A}{\partial W_i})^2 \sigma_{W_i}^2 + \sum_j (\frac{\partial A}{\partial W_j})^2 \sigma_{W_j}^2 + \sum_k (\frac{\partial A}{\partial W_k})^2 \sigma_{W_k}^2 \\ &+ \sum_i (\frac{\partial A}{\partial \phi_{rec,i}})^2 \sigma_{\phi_{rec,i}}^2 + \sum_j (\frac{\partial A}{\partial \phi_{rec,j}})^2 \sigma_{\phi_{rec,j}}^2 \end{aligned}$$

where we can calculate

$$\frac{\partial A}{\partial W_i} = \frac{\phi_{rec,i} b - \phi_{gen,i} a}{b^2} \qquad \frac{\partial A}{\partial W_j} = \phi_{rec,j} \frac{1}{b} \qquad \frac{\partial A}{\partial W_k} = \phi_{gen,k} (\frac{-a}{b^2})$$

$$\frac{\partial A}{\partial \phi_{rec,i}} = W_i \frac{b-a}{b^2} \qquad \frac{\partial A}{\partial \phi_{rec,j}} = W_j \frac{1}{b}$$

Finally we get

$$\begin{aligned} \sigma_A^2 &= \sum_i (\frac{\phi_{rec,i} b - \phi_{gen,i} a}{b^2} W_i)^2 + \frac{1}{b^2} \sum_j (\phi_{rec,j} W_j)^2 + (\frac{-a}{b^2})^2 \sum_k (\phi_{gen,k} W_k)^2 \\ &+ (\frac{b-a}{b^2})^2 \sum_i (W_i \sigma_{\phi_{rec,i}})^2 + \frac{1}{b^2} \sum_j (W_j \sigma_{\phi_{rec,j}})^2 \end{aligned} \qquad (7)$$

The several flux factors $\phi_{rec,i}, \phi_{rec,j}$ and $\phi_{gen,i}, \phi_{gen,k}$ are completely determine by the change of reference system and the $\sigma_{\phi_{rec,i}}, \sigma_{\phi_{rec,j}}$ are related to the resolution we can achieve in the reconstruction of the kinematic variables. For example if $\phi_{rec,i}, \phi_{rec,j}$ can be written as two functions of two kinematic variables that we note X and Y we have

$$\phi_{rec,i} = f(X_i, Y_i)$$

$$\phi_{rec,j} = f(X_j, Y_j)$$

We can generally assume that X and Y follow a normal law of variances σ_X and σ_Y, this variables can be choosen independant. Then $\sigma_{\phi_{rec,i}}, \sigma_{\phi_{rec,j}}$ can be expressed in terms of the resolutions of the kinematic variables

$$\sigma^2{}_{\phi_{rec,i}} = (\frac{\partial \phi_{rec,i}}{\partial X_i})^2 \sigma_{X_i}{}^2 + (\frac{\partial \phi_{rec,i}}{\partial Y_i})^2 \sigma_{Y_i}{}^2$$

$$\sigma^2{}_{\phi_{rec,j}} = (\frac{\partial \phi_{rec,j}}{\partial X_j})^2 \sigma_{X_j}{}^2 + (\frac{\partial \phi_{rec,j}}{\partial Y_j})^2 \sigma_{Y_j}{}^2$$

All the terms of the form $\frac{\partial \phi_{rec,j}}{\partial X_j}$ are determined by the change of the reference system, then if we know the variances σ_X and σ_Y we can completely determined the variance of the acceptance σ_A from formula (7).

5 Purity Computation

Keeping the same notations as in section (3) we define the purity

$$\mathcal{P} = \frac{N_{stay}}{N_{stay} + N_{leave}} = \frac{\sum\limits_i W_i}{\sum\limits_i W_i + \sum\limits_k W_k} = \frac{a}{b} \qquad (8)$$

where $\sum\limits_i W_i = N_{stay}$, $\sum\limits_k W_k = N_{leave}$ are independant variables, we can then derive the error calculation using the following formulae

$$\frac{\partial\mathcal{P}}{\partial W_i} = \frac{b-a}{b^2} \qquad \frac{\partial\mathcal{P}}{\partial W_k} = \frac{-a}{b^2}$$

We get for ${\sigma_{\mathcal{P}}}^2$

$${\sigma_{\mathcal{P}}}^2 = (\frac{b-a}{b^2})^2 \sum_i {W_i}^2 + (\frac{-a}{b^2})^2 \sum_k {W_k}^2$$

6 Efficiency Computation

6.1 general case

In a sample of events we can compute the efficiency of one or several cuts used to select one part of this sample. For this we evaluate in a given bin the ratio of the number of events which pass the cut over the total number of events in our sample. We write

$$\epsilon = \frac{\sum\limits_i W_i}{\sum\limits_p W_p} = \frac{a}{b}$$

$$\epsilon = \frac{N_{pass}}{N_{pass} + N_{cut}} = \frac{\sum\limits_i W_i}{\sum\limits_i W_i + \sum\limits_j W_j} \qquad (9)$$

where we note (i) the events which pass the cut and (j) the events which are cut, such that $\sum_i W_i = N_{pass}$ and $\sum_j W_j = N_{cut}$. These two variables are independant and the formula (1) can be applied for them. Then we can calculate the error on the efficiency

$$\sigma_\epsilon^2 = \sum_i (\frac{\partial\epsilon}{\partial W_i})^2 \sigma_{W_i}^2 + \sum_j (\frac{\partial\epsilon}{\partial W_j})^2 \sigma_{W_j}^2$$

where the derivatives can be written

$$\frac{\partial\epsilon}{\partial W_i} = \frac{b-a}{b^2} \qquad \frac{\partial\epsilon}{\partial W_j} = \frac{a}{b^2}$$

$$\sigma_\epsilon^2 = (\frac{(b-a)^2}{b^4}) \sum_i {W_i}^2 + \frac{a^2}{b^4} \sum_j {W_j}^2 \tag{10}$$

Finally we have

$$\sigma_\epsilon^2 = (\frac{b-2a}{b^3}) \sum_i {W_i}^2 + \frac{a^2}{b^4} \sum_p {W_p}^2 \tag{11}$$

6.2 Unweighted Case

In the unweighted case, we have $W_\xi = 1.$ with $\xi \in [i,j]$, thus if we note $n = \sum_i W_i$ and $N = \sum_p W_p$ we can write

$$\epsilon = \frac{n}{N}$$

$$\sigma_\epsilon^2 = \frac{n(N-n)}{N^3}$$

7 Comments on χ^2 computation

For N random variables X_i of mean values $\bar{X}_i$ and of variance σ_i we define

$$\chi^2 = \sum_{i=1}^{N} \frac{(X_i - \bar{X}_i)^2}{{\sigma_i}^2}$$

If the variables X_i follow the same probability distribution for example a normal law of mean $\bar{X}_i$ and variance σ_i that we note

$$f(X_i) = \mathcal{N}(X_i - \bar{X}_i, \sigma_i)$$

we can easily verify that the χ^2 follows for large N a probability distribution of the same type [2]

$$\mathcal{F}(\chi^2) \sim \mathcal{N}(\chi^2 - N, \sqrt{2N})$$

We can then compute the probability for the χ^2 to be greater than a given χ_0^2

$$Pr(\chi^2 \geq \chi_0^2) = \int_{\chi_0^2}^{\infty} F(\chi^2) d\chi^2$$

We define the confidence level : $1 - Pr(\chi^2 \geq \chi_0^2)$ which can be used to evaluate the distributions of the variables X_i. It means that if we don't know the distributions for the X_i we can make an hypothesis for them and then if the confidence level $1 - Pr(\chi^2 \geq \chi_0^2)$ calculated with this hypothesis is too low, our hypothesis has to be rejected. We can follow this procedure to find appropriate distributions for the X_i after a few iterations.

8 Conclusion

We have given a few formulae for the statistical treatement for important quantities such as efficiencies, acceptance and purity as well as for the prob-

[2] We have for the probability distribution of the N independant variables

$$f(x_1, ..., x_N) = \prod_{i=1}^{N} f(x_i) = \frac{1}{(2\pi)^{\frac{N}{2}} \prod_{i=1}^{N} \sigma_i} exp(-\sum_{i=1}^{N} \frac{(X_i - \bar{X}_i)^2}{{\sigma_i}^2})$$

Then we can express the probability distribution of the χ^2

$$\mathcal{F}(\chi^2) = \frac{1}{(2\pi)^{\frac{N}{2}} \prod_{i=1}^{N} \sigma_i} \int exp(-\sum_{i=1}^{N} \frac{(X_i - \bar{X}_i)^2}{{\sigma_i}^2}) \delta(\chi^2 - \sum_{i=1}^{N} \frac{(X_i - \bar{X}_i)^2}{{\sigma_i}^2}) \prod_{i=1}^{N} dx_i$$

A few calculations lead to a normal law for $\mathcal{F}$ with mean value N and variance $\sqrt{2N}$

$$\mathcal{F}(\chi^2) \sim \mathcal{N}(\chi^2 - N, \sqrt{2N})$$

ability distributions of a set of random variables. These calculations suppose an appropriate use of different theorems of statistics, we have shown on these few examples how to deal with them. The results of this contribution are useful in particular when we want to find out the statistical and numerical consequences of the weighting of a Monte Carlo in error computations. Applications can be found for examples in analyses of the proton structure functions and related subjects in the H1 experiment.

9 Acknowledgments

We are very thankful to Georges Cozzika for discussions and reading the manuscript.

References

[1] *Statistical Methods in Experimental Physics*, W.T. Eadie, D. Drijard, F.E. James et al. pub. North-Holland.

[2] *Techniques for Nuclear and Particle Physics Experiments*, W.R. Leo, pub. Springer-Verlag.